Université de Montréal

Statistical potentials for evolutionary studies

par
Claudia L. Kleinman

Département de Biochimie
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en Bioinformatique

June, 2010

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Statistical potentials for evolutionary studies

présentée par:

Claudia L. Kleinman

a été évaluée par un jury composé des personnes suivantes:

Sébastien Lemieux,	président-rapporteur
Hervé Philippe,	directeur de recherche
Nicolas Lartillot,	codirecteur
Mathieu Blanchette,	membre du jury
Gustavo Parisi,	examinateur externe
Joelle Pelletier,	représentant du doyen de la FES

Thèse acceptée le: . . . . . .

RÉSUMÉ

Les séquences protéiques naturelles sont le résultat net de l'interaction entre les mécanismes de mutation, de sélection naturelle et de dérive stochastique au cours des temps évolutifs. Les modèles probabilistes d'évolution moléculaire qui tiennent compte de ces différents facteurs ont été substantiellement améliorés au cours des dernières années. En particulier, ont été proposés des modèles incorporant explicitement la structure des protéines et les interdépendances entre sites, ainsi que les outils statistiques pour évaluer la performance de ces modèles. Toutefois, en dépit des avancées significatives dans cette direction, seules des représentations très simplifiées de la structure protéique ont été utilisées jusqu'à présent.

Dans ce contexte, le sujet général de cette thèse est la modélisation de la structure tridimensionnelle des protéines, en tenant compte des limitations pratiques imposées par l'utilisation de méthodes phylogénétiques très gourmandes en temps de calcul. Dans un premier temps, une méthode statistique générale est présentée, visant à optimiser les paramètres d'un potentiel statistique (qui est une pseudo-énergie mesurant la compatibilité séquence-structure). La forme fonctionnelle du potentiel est par la suite raffinée, en augmentant le niveau de détails dans la description structurale sans alourdir les coûts computationnels. Plusieurs éléments structuraux sont explorés : interactions entre pairs de résidus, accessibilité au solvant, conformation de la chaîne principale et flexibilité. Les potentiels sont ensuite inclus dans un modèle d'évolution et leur performance est évaluée en termes d'ajustement statistique à des données réelles, et contrastée avec des modèles d'évolution standards. Finalement, le nouveau modèle structurellement contraint ainsi obtenu est utilisé pour mieux comprendre les relations entre niveau d'expression des gènes et sélection et conservation de leur séquence protéique.

Mots clès : Évolution moléculaire, structure des protéines, Markov chain Monte Carlo, maximum de vraisemblance, statistique Bayesienne, potentiels statistiques.

ABSTRACT

Protein sequences are the net result of the interplay of mutation, natural selection and stochastic variation. Probabilistic models of molecular evolution accounting for these processes have been substantially improved over the last years. In particular, models that explicitly incorporate protein structure and site interdependencies have recently been developed, as well as statistical tools for assessing their performance. Despite major advances in this direction, only simple representations of protein structure have been used so far. In this context, the main theme of this dissertation has been the modeling of three-dimensional protein structure for evolutionary studies, taking into account the limitations imposed by computationally demanding phylogenetic methods. First, a general statistical framework for optimizing the parameters of a statistical potential (an energy-like scoring system for sequence-structure compatibility) is presented. The functional form of the potential is then refined, increasing the detail of structural description without inflating computational costs. Always at the residue-level, several structural elements are investigated: pairwise distance interactions, solvent accessibility, backbone conformation and flexibility of the residues. The potentials are then included into an evolutionary model and their performance is assessed in terms of model fit, compared to standard evolutionary models. Finally, this new structurally constrained phylogenetic model is used to better understand the selective forces behind the differences in conservation found in genes of very different expression levels.

Keywords: molecular evolution, protein structure, Markov chain Monte Carlo, maximum likelihood, Bayesian statistics, statistical potentials.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ADH	Alcohol dehydrogenase
DS	Data set
CV	Cross-validation
EM	Expectation maximization
KLD	Kullback-Leibler divergence
MCMC	Monte Carlo Markov chain
MJ	Miyazawa and Jernigan
ML	Maximum likelihood
MSA	Multiple sequence alignment
PDB	Protein Data Bank
SC	Structurally constrained

A mis hijos, como todo lo demás.

ACKNOWLEDGMENTS

As I write the last lines of this dissertation, I cannot help but thinking of all the people that made it possible. There are many, many people I am grateful to: some of them inspired me, some made these years worth remembering, and some of them helped me through the difficult times.

First, I want to thank my advisor Hervé Philippe for the many things I have learnt from him. The thoroughness when analyzing data, the respect for other's ideas, the *esprit critique*, and the constant search for a different angle to look at things. Also, for his patience and support while I was struggling to balance research with family life.

I thank my co-supervisor, Nicolas Lartillot, for his intelligent ideas. Without him, this project would not have even started, let alone succeeded in any way. And, of course, for the opportunity to visit the beautiful France. This work would not have been possible without Nicolas Rodrigue either, who was involved in it from the very first day. I thank him for the knowledge, suggestions and hours of work he invested in this project.

I am especially grateful to Gertraud Burger, who allowed me to land here from the other end of the world. For the hope she had in me even before I could prove anything, and for always showing me nothing but honesty, care and respect.

My sincere thanks to all the people in the Cedergren Bioinformatics Center, who made it such an enriching environment. To Henner Brinkmann, for his help assembling datasets, his cocktails and the countless coffee breaks. To Jean-Christophe Grénier, who was a pleasure to work with, for his contribution to the project as an internship student. To Franz Lang, for his clever and always troubling questions. To Béatrice Roure, Yao-qing Shen, Sivakumar Kannan, Dorothée Coste, Pasha Javadi, Véronique Marie, Fabrice Baro, Shona Tejeiro, Lise Forget and Natacha Beck, for their friendship and support.

The years spent in Montreal would not have been as wonderful without my friends Joannie Roy, Rocío González-Lamothe, Naiara Rodríguez-Ezpeleta, Olivier Jeffroy, Darío Kunik and Marie-ka Tilak. Thanks to them, I have found my home here.

I would like to express my gratitude to Elaine Meunier, Marie Robichaud and Marie Pageau. Their efficiency and kindness made university life much, much easier.

I am also grateful to José María Delfino for welcoming me twice in his lab, for many scientific discussions and invaluable suggestions. And to Javier Santos, whose joy and enthousiasm for research are contagious.

During the development of this dissertation I received financial support from the Natural Sciences and Engineering Research Council of Canada, the biT fellowships for excellence (a Canadian Institutes of Health Research strategic training program grant in bioinformatics), the Bank of Montreal, the Université de Montréal and the Québec Education Ministry. I also acknowledge the Réseau Québécois de Calcul de Haute Performance for the computational ressources they provided.

Finally, I am forever indebted to my family. My parents Graciela and Hugo are, of course, ultimately responsible for everything good that came out me. Words of thanks cannot begin to express the gratitude I feel towards them; I hope I made them proud. Thanks to Tito and Nora, for their faith in me and their constant words of encouragement; they meant a lot, each and every one of them. To my three sisters, Popi, Andru and Mailen, who are closer than they think, always on my mind, always in my heart. To my grandmother Chiche, an example of a strong, independent, professional woman. And to Patricia, Roberto, Juan and Angela.

No one, however, helped me more constantly and directly in pursuing this path than Santiago, without whom I would have been lost. During all these years, he has been my Ariadne's thread, my support and motivation, my guide and inspiration, my everything. Thank you, Doctor.

AT A GLANCE

Just as an archeologist studying ancient buildings to understand human culture, we will look at three-dimensional protein structures to search for clues on the evolutionary processes that shaped them. Imagine for a moment that you find the remains of an unknown civilization. You can learn many things just by looking at the construction materials, from the approximate time this civilization lived to many environmental and social conditions. For example, the fragile materials used in precarious houses of a brazilian *favela* are impossible to find in a northern canadian home: the rigorous winter weather makes them unfit for survival. Branches, mud and palm leaves belong to a tropical place, and to a certain social class. Big heavy stones and marbles, in turn, point to a complex social organization with division of labour, capable of sustaining constructions over long periods of time.

An archeologist, however, would not stop at the analysis of raw materials. The particular way a construction is made, its organization, its level of complexity, the presence of elements imported from other cultures, all the pieces are essential to the puzzle. When taken together, they speak of social relations, economic systems (nomad or sedentary, agricultural or industrial, producer or consumer). In the same way, we will turn to the buildings in a cell -proteins- to incorporate what we know about them and improve our understanding of the underlying evolutionary process.

But first, we need to determine the elements we will be focusing on. To what level of detail is it worth going into? When are the main traits, such as size, stability, durability of a building enough to draw conclusions, and when is it worth describing detailed features like ornaments and colors? Prisons, hospitals and schools need a particular internal organization, as many enzymes do. For monuments and churches, the exterior counts just as much, as it does in binding proteins. Sometimes, a robust construction is essential: like skyscrapers in earthquake zones, proteins in thermophilic organisms will need an

above-average stability. In other cases, aesthetics and design are the dominant traits. Can we define some general features that are important in all the cases?

The main theme of my PhD work has been the modeling of three-dimensional protein structure in a meaningful way for an evolutionary perspective, while taking into account the limitations imposed by computationally highly demanding phylogenetic methods.

The first chapter presents a general introduction to the concepts from structural and evolutionary biology that are merged in the rest of the work. First, the statistical modeling of molecular evolution is introduced, focusing on the description of selective constraints, in particular those imposed by the protein structure. Then, a discussion on the available tools for evaluating the sequence-structure compatibility follows, from first principles to knowledge-based methods (or statistical potentials), which we chose for our approach.

The second chapter introduces an optimization method of statistical potentials conceived for evolutionary studies. In evolution, a protein's structure changes very slowly, while a multitude of sequences generated by random mutations have to conform to this structure. We thus posed the problem in terms of protein design, or the inverse folding problem, that is, predicting sequences compatible with a given structure. A probabilistic formulation is developed, where the goal is to obtain a probability distribution of sequences conditional on a structure. An alternative optimization procedure, allowing to considerably decrease the computational time required, is presented in Appendix 1.

In the third chapter, the functional form of the statistical potential is refined, adding several structural elements to the three dimensional description of the proteins and studying their impact on the evolution of sequences. To do so, the new potentials are included into a structurally constrained phylogenetic model, and their statistical fit to real data is assessed in a Bayesian framework.

Next, chapter four presents a series of results on the application of this framework to a relatively larger set of proteins, to study the variability of the influence of protein structure on sequence evolution. In particular, we examine how this influence is modulated

by transcriptional properties of the encoding genes, by contrasting patterns of model fit obtained on proteins of different expression level.

Finally, chapter five presents concluding remarks and future directions.

The work presented in this dissertation is previously published in the following articles:

Chapter 2:

C. L. Kleinman, N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot. A maximum likelihood framework for protein design. *BMC Bioinformatics*, 7 :326, 2006.

Appendix 1:

C. Bonnard, C. L. Kleinman, N. Rodrigue, and N. Lartillot. Fast optimization of statistical potentials for structurally constrained phylogenetic models. *BMC Evolutionary Biology*, 9(1) :227, 2009.

Chapter 3:

C. L. Kleinman, N. Rodrigue, Nicolas Lartillot, and H. Philippe. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol*, Feb 16. [Epub ahead of print], 2010.

The computational methods for including the potentials into a phylogenetic framework and evaluating their model fit are not formally included in this thesis, but published in the article:

N. Rodrigue, C. L. Kleinman, H. Philippe, and N. Lartillot. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol*, 26(7) :1663-76, 2009.

Finally, the structural analysis performed in the following article was also used to gain insights for the work presented here, but was not included in its totality in this dissertation:

J. Santos, C. Marino-Buslje, C. L. Kleinman, M R. Ermácora, and J.M. Delfino. Consolidation of the thioredoxin fold by peptide recognition: Interaction between *E. coli* thioredoxin fragments 1-93 and 94-108. *Biochemistry*, 46(17) :5148Ð5159, 2007.

CHAPTER 1

INTRODUCTION

1.1 Modeling evolution at the molecular level

Traditionally restricted to biological classification and descriptive reconstructions of species history, phylogenetic studies have now pervaded almost every discipline in biological sciences; in any comparison of related sequences, there is a phylogeny implicitly assumed. Comparative sequence analysis is routinely used for a range of diverse applications, from the identification of functional regions in genomes (Margulies and Birney, 2008) to structural homology modeling (Madhusudhan et al., 2005). As the amount of public biological data increases, so does the need for a deeper understanding of the dependencies and patterns that originated from a shared evolutionary history.

Studying molecular evolution presents, however, unique challenges. We seek to infer the evolutionary scenario most consistent with the incomplete information contained in the data, usually limited to alignments of contemporary sequences; there is virtually no direct information about the past. Furthermore, what makes evolutionary studies particularly difficult (and thus, interesting) is that these observed sequences, as we understand it now, are the net outcome of the interplay of mutation, natural selection and stochastic variation due to genetic drift, with each one of these processes adding a layer of complexity to the problem. Evolution of protein sequences is not determined exclusively by selection on protein structure and function, but is also affected by a panoply of diverse, complex, overlapping and sometimes contradictory factors. For example, restricting ourselves to gene features defined at the cellular level and that have been correlated to the rate of protein evolution (reviewed in Pal et al., 2006), we find variations in mutation (Ellegren et al., 2003; Lercher et al., 2001) and recombination (Lercher and Hurst, 2002; Betancourt and Presgraves, 2002) rates associated with genomic posi-

tion, gene dispensability (Hirsh and Fraser, 2001; King Jordan et al., 2002; Wall et al., 2005), protein structure and stability (Haney et al., 1999; Goldman et al., 1998; Dean et al., 2002), position in biological networks (Aris-Brosou, 2005; Fraser et al., 2002) and transcriptional properties, such as expression breadth and expression level (Duret and Mouchiroud, 2000; Subramanian and Kumar, 2004; Wright et al., 2004; Drummond et al., 2006; Drummond and Wilke, 2008).

A great amount of biological knowledge has been built on many of these elements over the last 30 years, which could be potentially included in a comprehensive view of protein evolution. Always with the goal of forming an integral view of protein evolution, one strategy is to tackle one particular aspect at a time. In the present work, we will try to include as much previous knowledge as possible, but only on a subset of all the selective constraints: the ones imposed by the requirements of maintaining a three-dimensional protein structure.

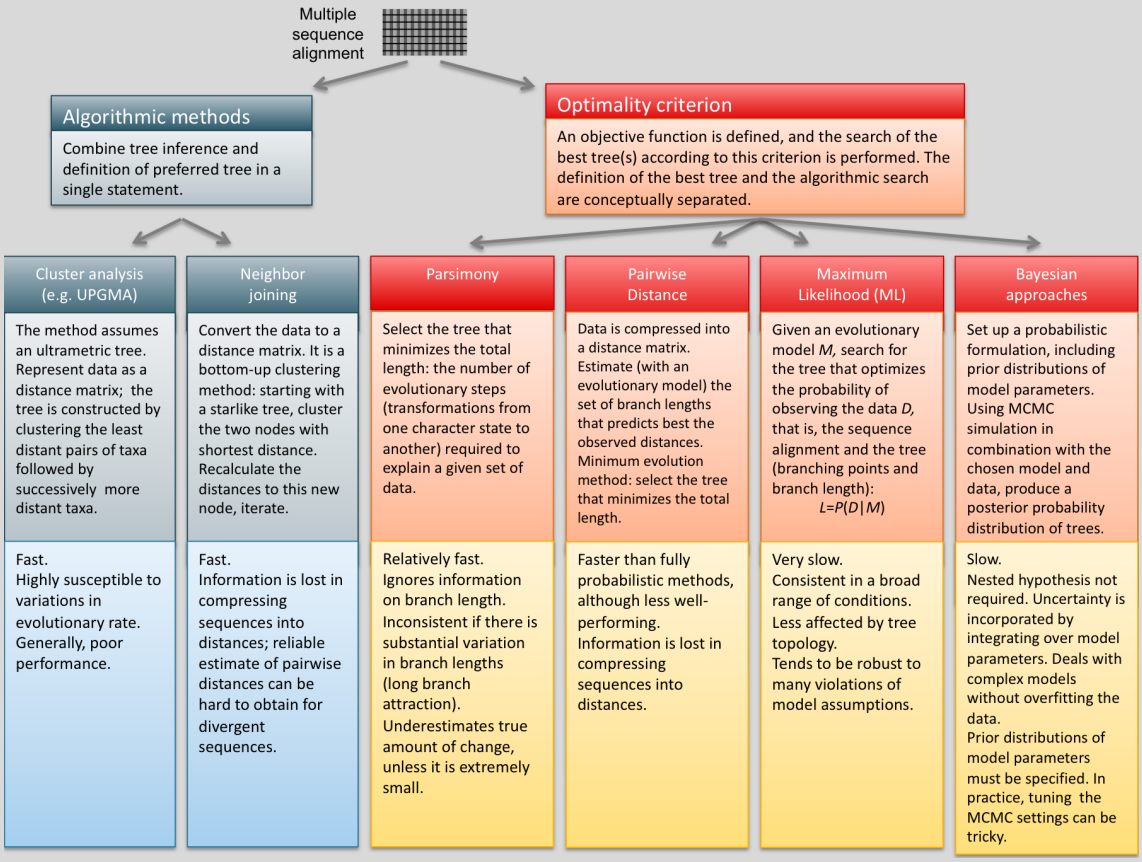
For the challenges described above, probabilistic methods based on explicit models of evolutionary change, along with statistical tools for hypothesis testing, are of particular interest in this field, where traditional evaluation of hypotheses by experimental procedures is almost never an option. For the sake of brevity, and given the broad scope of the subject area, I will focus the discussion on probabilistic methods, and omit the explanation of the alternative parsimony, distance-based and algorithmic approaches to phylogenetic inference. A brief outline is nonetheless presented in Box 1.1; for a thorough review, see Swofford et al. (1996) and Felsenstein (2004).

Compared to the alternatives, probabilistic methods present several advantages in our context. The set of assumptions they rely on is made entirely explicit through a *model*. The laws of probability provide a guarantee that the available empirical evidence (the data) has been analyzed in the framework of this model in a logically coherent fashion. Model violations, given this explicit statement of assumptions, are easier to interpret and evaluate. Finally, model selection theory has a long history in the statistical literature, and several methods have been adapted to the evolutionary problem (Sullivan and Joyce,

2005).

Probabilistic approaches to phylogenetic inference come in two flavors: maximum likelihood and Bayesian approaches. Bayesian approaches of model comparison, which have been extensively developed during the last few years (Box 1.2), are attractive for two main reasons. First, they dispense with the need of analytical integrations by the use of simulation-based numerical strategies as Markov chain Monte Carlo (MCMC). And second, they allow the evaluation of more complex models, since they implicitly penalize overly high-dimensional parameterizations by integrating away nuisance parameters (Gelman et al., 2004; Huelsenbeck et al., 2002).

Box 1.1: Overview of phylogenetic inference methods.



Generally, sequence evolution is described using two components: a phylogenetic tree and a mathematical description of the way individual sequences evolve by nucleotide or amino acid replacement along the branches of that tree (Swofford et al., 1996; Lio and Goldman, 1998; Whelan et al., 2001). These replacements are considered as the product of chance substitution events, and their occurrence at each site is mathematically modeled by a Markov process: a stochastic process with a finite number of possible states -the sequence characters at each site- and some known probabilities p_{ij} of moving from state i to state j on a given time duration. The probability of change of one sequence into another is dependent only on the current state of the system, and not on its previous history; in other words, it is a memoryless process. Defined in this way, the whole process is entirely specified by the matrix of transition probabilities p_{ij}, which thus takes a central place in this framework. The development of more accurate and realistic models of sequence evolution has received much attention in recent years, in hopes of reducing phylogenetic reconstruction artifacts due to model misspecifications (Philippe et al., 2005; Lartillot et al., 2007), as well as of addressing particular aspects of molecular evolution (Pal et al., 2006), as I will do in this dissertation.

Probabilistic methods in phylogenetics thus evaluate a hypothesis about evolutionary history in terms of the likelihood (i.e. the probability) that a proposed model and the hypothesized history would produce the observed data. When phylogeny is the problem of interest, it is then inferred by finding those trees that yield the highest likelihood (Box 1.1). Alternatively, the object of study may be the evolutionary process itself, with the model of sequence change as the hypothesis under scrutiny. In this case, although a joint estimation is possible, a given species' phylogeny is considered as known to gain insights into the mechanisms of molecular evolution. The observed data thus includes a sequence alignment and the corresponding tree topology, and the probability of a proposed evolutionary model given this data is evaluated.

Until recently, the computational cost of probabilistic phylogenetic methods placed severe practical limits on the complexity of the problems that could be handled, forcing

the use of simplifying assumptions that are not always biologically reasonable. This has dramatically changed in recent years with the large increase in computing power, and the parallel development of simulation-based numerical integration strategies like Markov chain Monte Carlo (MCMC). All these tools allow the simultaneous comparison of different models, setting the grounds for an iterative research cycle where evolutionary models are progressively refined and contrasted against each other. A greater level of realism is conceivable, dropping some of the simplifications that have been made in the modeling of evolution, among which the independence between sites and the omission of the three dimensional structure of proteins. Before getting into the details, let us overview the most commonly used models of molecular evolution.

Box 1.2: Bayesian model comparison.
Bayesian inference is based on the analysis of the posterior probability distribution over the parameters of interest. Given a model M, with a parameter vector $\theta \in \Theta$ (specifying, for instance, the tree topology and branch lengths, or the parameters of the substitution model, see section 1.1.1), and applied on a dataset D, the posterior probability distribution is given by Bayes' theorem:

$$p(\theta \mid D,M) = \frac{p(D \mid \theta,M)p(\theta \mid M)}{p(D \mid M)} \tag{1.1}$$

where $p(\theta \mid M)$ is the prior distribution, $p(D \mid \theta,M)$ the likelihood function and

$$p(D \mid M) = \int_{\Theta} p(D \mid \theta,M)p(\theta \mid M)d\theta \tag{1.2}$$

is a normalization constant, also called the *predictive probability* or *marginal likelihood*. Parameter estimation is done by computing expectations over the posterior distribution of equation 1.1. In particular,

$$\bar{\theta} = \int \theta p(\theta \mid D,M)d\theta \tag{1.3}$$

The analytical calculation of these high dimensional integrals is often not feasible. One way to numerically solve them is to use a Markov chain Monte Carlo walking to sample θ, using the mean of this sample to approximate expectations.
When performing statistical comparisons, the marginal likelihood (1.2) is of primary importance. As a function of M, it can be directly interpreted as the likelihood of the model M, given that we observe the data D. The preferred model will be thus the one of greatest marginal likelihood. When two particular models M_1 and M_2 are being compared, the *Bayes factor* in favor of M_1 over M_2 is defined as the ratio of their respective marginal likelihoods (Jeffreys, 1935; Kass and Raftery, 1995):

$$B_{01} = \frac{p(D \mid M_1)}{p(D \mid M_2)} \tag{1.4}$$

Values of Bayes factor greater than 1 will be considered as evidence in favor of M_1, and vice-versa. The numerical estimation of this value (Neal, 2000; Gelman, 1998; Lartillot and Philippe, 2006) is challenging but feasible; several approximation strategies have been explored for alleviating the computational cost of this calculation (Rodrigue et al., 2007, 2009).
Other approaches for evaluating model adequacy in a Bayesian context are available. In particular, posterior predictive checking has been proposed (Rubin, 1984; Gelman et al., 1996; Rodrigue et al., 2006), where discrepancies between features of true data and data simulated under the model of interest are analyzed. In the work presented in this dissertation, however, we have focused on the traditional Bayes factor, which offers a very intuitive interpretation in a model-selection perspective.

1.1.1 Models of molecular evolution

Setting the grounds: DNA models

Models of molecular evolution will be presented in a logical progression, rather than in chronological order of appearance. The simplest models describe evolution as a homogeneous stochastic process that acts on DNA sequences, by accumulating substitutions according to a matrix of rates (substitutions per site per unit of evolutionary time) at which a nucleotide is replaced by an alternative nucleotide. By assuming that sites evolve independently, a single character (position) can be considered in isolation from the rest. Further assumptions include time-reversibility and stationarity (that is, a process at equilibrium). Many of these assumptions, necessary at first for rendering the study of molecular evolution into a mathematically tractable form, have been (and continue to be) relaxed, as we will see later.

In its most simple form, the instantaneous rate matrix can be specified with a single free parameter, as is the case in the Jukes-Cantor model (Jukes and Cantor, 1969):

$$Q_{JC} = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix} \tag{1.5}$$

where the rows (and columns) correspond to the bases A, C, G and T, respectively, and the (i, j) entry represents the rate at which a base i is replaced by a base j. The process is homogeneous in every possible sense: not only are the rates all equal, but also the matrix remains constant over time and for different sites in the alignment.

Transition probabilities $P_{ab}(t)$ from site a to site b over time $t > 0$ can be derived from the instantaneous rate matrix Q. The entries of Q specify the rate of change from state a to state b in an infinitesimal time interval, with diagonal entries chosen so that

the sum of the row elements equals zero (i.e. $Q_{aa} = -\sum_{b \neq a} Q_{ab}$). $P(t)$ can be derived solving the differential equation $\mathrm{d}P(t)/\mathrm{d}t = P(t)Q$. The solution to this equation, given the start assumption $P(0) = I$, is the following:

$$P(t) = e^{tQ} = \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!}, \tag{1.6}$$

which can be solved through diagonalization of Q (Lio and Goldman, 1998). Assuming that every site evolves independently, the likelihood of each site is calculated via the pruning algorithm (Felsenstein, 1981), and the full likelihood is subsequently computed by taking the product of each individual site likelihood over all sites in the alignment. The pruning algorithm requires a computational time proportional to the sequence length N, the number of taxa and the square of the number of characters states m allowed at each site (for nucleotide models, $m = 4$).

The limitations of the extremely simple model in (1.5) are readily apparent, and it has been expanded substantially. Kimura (1980) proposed a two-parameter model that distinguishes between transition and transversion rates; further related contributions consist in considering asymmetries between some of the reciprocal changes (Blaisdell, 1985), a four-parameter (Takahata and Kimura, 1981) and a six-parameter model (Gojobori et al., 1982). Felsenstein (1981) proposed a model in which the rate of substitution to a nucleotide depends on the equilibrium frequency of that nucleotide, accounting thus for the nucleotide base composition heterogeneity in DNA sequence data:

$$Q_{F81} = \begin{bmatrix} \bullet & \mu\pi_T & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \bullet & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \bullet & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \mu\pi_C & \bullet \end{bmatrix} \tag{1.7}$$

For clarity, and since the rows of the matrix are constrained to sum zero, I have used

dots in the diagonal, and will do so in the subsequent matrices. Hasegawa et al. (1985) combined this model with Kimura's model by accounting for transition/transversion bias:

$$Q_{HKY85} = \begin{bmatrix} \bullet & \beta\pi_T & \beta\pi_C & \alpha\pi_G \\ \beta\pi_A & \bullet & \alpha\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha\pi_T & \bullet & \beta\pi_G \\ \alpha\pi_A & \beta\pi_T & \beta\pi_C & \bullet \end{bmatrix} \tag{1.8}$$

Once again, the rate of exchange between each nucleotide can be further parameterized. It is generally convenient to decompose Q into two matrices, (ρ) and (π), representing the exchangeability parameters and the equilibrium frequencies, respectively. In the most general case of these kind of DNA models, these matrices are:

$$(\rho) = \begin{bmatrix} \bullet & \rho_{AC} & \rho_{AG} & \rho_{AT} \\ \rho_{CA} & \bullet & \rho_{CG} & \rho_{CT} \\ \rho_{GA} & \rho_{GC} & \bullet & \rho_{GT} \\ \rho_{TA} & \rho_{TC} & \rho_{TG} & \bullet \end{bmatrix} \tag{1.9}$$

and

$$(\pi) = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix} \tag{1.10}$$

where $\sum_{1 \leq m \leq 4} \pi_m = 1$. The off-diagonal elements of Q are then equal to the off-diagonal elements of the matrix product of ρ and π, and the diagonal elements of Q

are once again set such as the sum of the elements of each row equals zero. Constraining Q into a symmetric matrix such that $\rho_{lm} = \rho_{ml}$ ensures time-reversibility and yields the GTR (General time-reversible) model. This is the model that will be used later on as a basis (describing the mutational process) for constructing a mechanistic model of the mutation-selection process that operates on the evolution of protein sequences. Although this parameterization does not reflect tendencies for mutation rates to be context-dependent, it constitutes a good compromise between accuracy in the description of the mutational process and conceptual simplicity, important in this stage of model development. More realistic treatments of the mutation process could be added in a later stage.

Considering protein phenotype: amino acid models

A first direction to more explicitly incorporate selective effects on the substitution process of protein genes is to model evolution directly at the amino acid level, where the phenotypic expression of a genetic change is most evident. Substitutions that change the amino acid sequence should have a more drastic effect on phenotype, and a lower probability of fixation for molecules under purifying selection. A large body of empirical evidence shows that the rate of exchange between different amino acids is affected by the physicochemical characteristics of the amino acids involved, and this effect is very difficult to capture with a DNA model that treats all sites as equal, without regard to the amino acid sequence encoded.

Formally, amino acid models of sequence evolution are similar to the nucleotide models described so far, but with a state space of 20 characters instead of 4. In contrast with DNA substitution models, however, empirical matrices are generally used, which at first have been obtained by counting pairs of amino acids at homologous positions in large sets of aligned proteins (Dayhoff et al., 1978; Jones et al., 1992b). More recently, matrices optimized by maximum likelihood have also been proposed for mitochondrial (Adachi, 1996), chloroplast (Adachi et al., 2000), and nuclear (Whelan and Goldman, 2001) proteins. Typically, a hybrid model is used in the phylogenetic analysis of amino

acid sequences: equilibrium frequencies are estimated from the analyzed data, while the exchangeability parameters are taken from one of the empirical models described above.

Including the genetic code: codon models

The effects of effectively changing the amino acid sequence by a site substitution are better captured by approaches based on amino acid models than with a nucleotide-based model. However, the former have the drawback of not taking into account the genetic code structure and of confounding mutation and selection, by only analyzing the net effect of these two complex processes. A third class of models, formulated at the codon level, circumvent these two limitations.

In order to accommodate the structure of the genetic code, the definition of a *site* (that is, the unit of substitution) is changed from single nucleotide to triplets of nucleotides. A distinction is made between substitutions that change the encoded amino acid (nonsynonymous) and the ones that do not (synonymous). Assuming that synonymous substitutions are neutral, codon evolution is thus described as a combination of changes at the nucleotide level and selective constraints operating at the protein level.

The most widely used codon models are modifications of those originally proposed by Muse and Gaut (1994) and Goldman and Yang (1994). In both cases, instantaneous changes at more than one codon position are disallowed, as well as changes to premature stop codons. The state space of these Markov models is thus increased to the 61 sense codons of the universal genetic code. The original formulation of Muse and Gaut (MG) has a rate matrix of the form

$$Q_{MG_{ab}} = \begin{cases} 0 & \text{if } a \text{ and } b \text{ differ by more than one codon position} \\ \alpha\pi_b & \text{synonymous substitution} \\ \beta\pi_b & \text{nonsynonymous substitution} \end{cases} \tag{1.11}$$

The parameter π_b, representing the equilibrium frequency of the nucleotide type in the target codon, accounts for compositional heterogeneity at the nucleotide level. At the phenotypic level, this model allows for a different rate of substitution for synonymous (α) and nonsynonymous (β) events.

In the Goldman and Yang (GY) model, the equilibrium frequency of the target codon (as opposed to nucleotide type) is used to describe constraints at the DNA level (noted here $\pi_{c(b)}$). A parameter accounting for transition/transversion bias (κ) is included. To account for selective constraints at the amino acid level, substitution rates are modified by a multiplicative factor in the case of a nonsynonymous event. The definition of these factors is based on an amino acid distance matrix, derived by comparing physicochemical properties of the 20 amino acids (Grantham, 1974). The off-diagonal elements of the rate matrix are defined as follows:

$$Q_{GY_{ab}} = \begin{cases} 0 & \text{if } a \text{ and } b \text{ differ by more than one codon position} \\ \mu\pi_{c(b)}e^{-d_{AA_aAA_b}/V} & a \text{ and } b \text{ differ by a transversion} \\ \mu\kappa\pi_{c(b)}e^{-d_{AA_aAA_b}/V} & a \text{ and } b \text{ differ by a transition} \end{cases} \tag{1.12}$$

where $d_{AA_aAA_b}$ is the physicochemical distance between amino acids a and b, and V is a tuning parameter allowing the distance matrix to better fit the data, accounting for differing levels of sequence variability between genes. This model was later simplified by the authors to estimate selective pressure explicitly using the single parameter ω (Yang, 1998). This is essentially equivalent to the treatment in the MG model, setting $\alpha = \mu$ and $\beta = \mu\omega$. If frequent amino acid changes present a selective advantage, the nonsynonymous substitution rate will be higher than the synonymous rate, and as a result $\omega > 1$. Conversely, the case where purifying selection acts to preserve amino acid sequence corresponds to $\omega < 1$. Neutrally evolving sequences exhibit similar synony-

mous and nonsynonymous rates, and thus $\omega \approx 1$. The simplified version of this model is thus:

$$Q_{GY_{ab}} = \begin{cases} 0 & \text{if } a \text{ and } b \text{ differ by more than one codon position} \\ \mu\pi_{c(b)} & a \text{ and } b \text{ differ by a synonymous transversion} \\ \mu\pi_{c(b)}\kappa & a \text{ and } b \text{ differ by a synonymous transition} \\ \mu\pi_{c(b)}\omega & a \text{ and } b \text{ differ by a nonsynonymous transversion} \\ \mu\pi_{c(b)}\kappa\omega & a \text{ and } b \text{ differ by a nonsynonymous transition} \end{cases} \tag{1.13}$$

Codon models formulated in this way have direct connections to population genetic theory (Thorne et al., 2007; Yang and Nielsen, 2008). If the rate of a sequence change can be separated into factors corresponding to mutation (Q^{mut}) and to natural selection, then the factor associated with natural selection represents the probability of fixation. For a population of size $\aleph$, we have, at the population level:

$$Q_{ab} = 2\aleph Q^{mut}_{ab} p_{fix}(ab) \tag{1.14}$$

Since a neutral substitution in a diploid population has a probability of fixation

$$p^0_{fix}(ab) = \frac{1}{2\aleph} \tag{1.15}$$

We can write the rate of substitution from a to b as:

$$Q_{ab} = Q^{mut}\left(\frac{p_{fix}(ab)}{p^0_{fix}(ab)}\right) \tag{1.16}$$

In this way, the selective parameter ω of equation 1.13, as well as the term $e^{-d_{AA_aAA_b}/V}$ of equation 1.12 can be understood as a ratio of fixation probabilities. This is particu-

larly attractive for our purposes, that is, assessing selective constraints related to the protein structure. In addition to providing an improvement in model realism for protein coding sequences, codon models can also be designed to test hypotheses about the selective pressures operating on sequences (reviewed in Delport et al., 2009; Anisimova and Kosiol, 2009).

Although the codon models presented so far are mechanistically motivated, in the sense that translation of proteins is explicitly considered via the genetic code structure and the separation of processes acting at the DNA and amino acid level, only the net resultant of selection is captured in the parameters modulating nonsynonymous substitutions. Our motivation in the present work, however, is to focus exclusively in the constraints imposed by the three dimensional structure, in order to disentangle these structural constraints from other selective forces.

Increasing the state space in codon models produces high computational demands, which has prevented their widespread use for phylogenetic inference and the development of more complex versions of these first models. This has started to change in recent years, and several extensions relaxing some of the initial simplifying assumptions have been proposed (reviewed in Anisimova and Kosiol, 2009), allowing for the development of a mechanistic modeling alternative with explicit consideration of protein structure, as we will see in section 1.1.2.

1.1.2 Towards more realistic models: the case of tertiary structure

Proteins require a suitable three dimensional structure to function. Substitutions that affect the stability of the folded state will have a deleterious effect on fitness, and a lower probability of fixation. Stability of the native state is not, however, the only structural requirement of a viable molecule. For example, exposure of particular combination of amino acids on the surface, enabling the protein to interact inappropriately with a

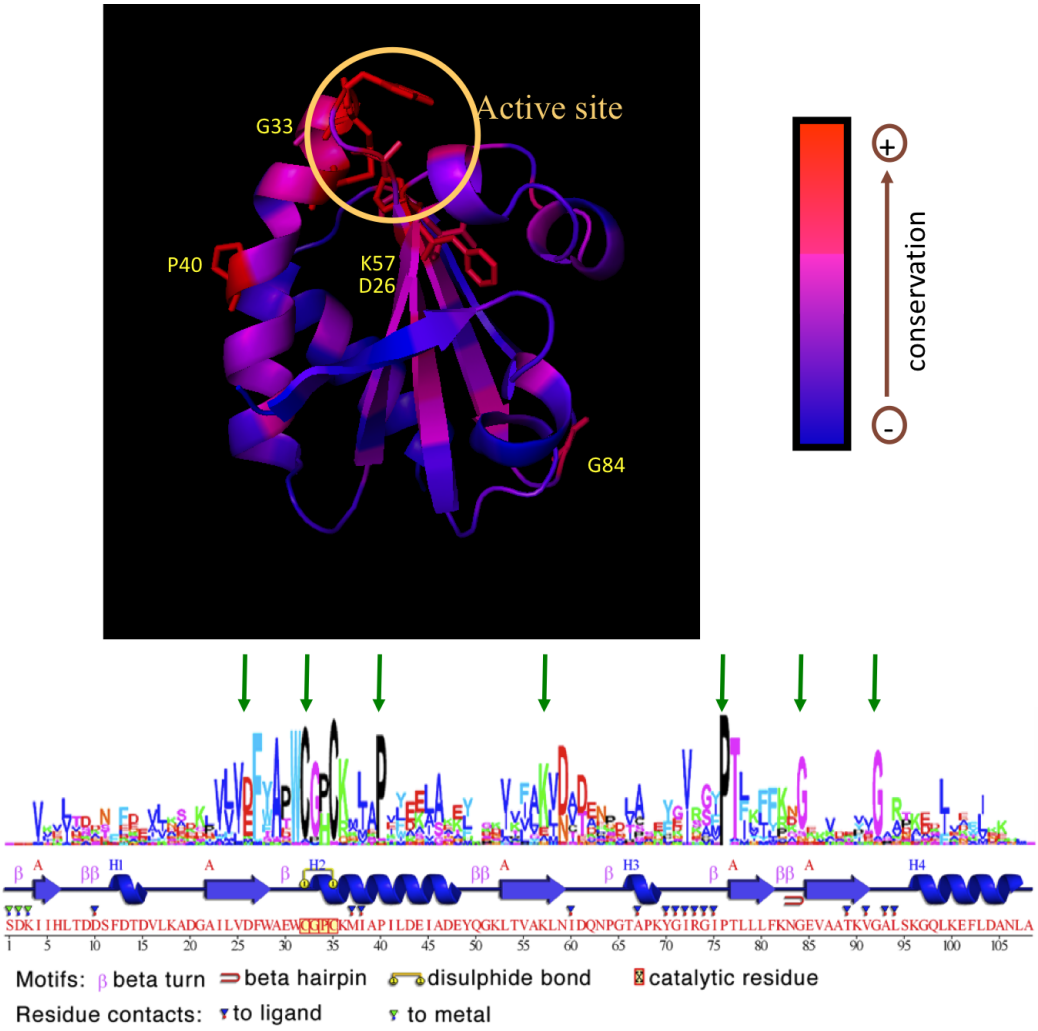


Figure 1.1: **Sequence conservation mapped onto the crystallographic structure of thioredoxin 2TRX.** Sequence profiles generated from a multiple sequence alignment (MSA) of 162 eukaryotic sequences, as described in chapter 3. The frequency of the 20 amino acids a at each position i was computed, yielding a vector $q_i(a)$ of site-specific profiles. In the graphic representation, the total height h_i at each position is proportional to the Shannon information as follows: $h_i = \sum_a q_i(a) ln q_i(a)$. These Shannon information scores were then mapped on the protein structure according to the color scheme on the right. Secondary structure representation taken from PDBsum (Laskowski, 2009). Bottom: native sequence of the reference structure, from *E. coli*.

wide range of cellular components may induce aggregation (Bucciantini et al., 2002; Dobson, 2003), causing unspecific cellular toxicity. In another example, flexibility and mobile regions are important for the function of an enormous number of proteins (see for example Gerstein and Echols, 2004; Wilson and Brunger, 2000; Huse and Kuriyan, 2002).

All these requirements produce heterogeneous substitution processes across sites that are evident at first sight in the alignments of homologous sequences; an illustration is presented in figure 1.1. Some of the positions owe their level of conservation to constraints related to the specific function of the protein (e.g. active site of the enzyme or ligand binding sites). Some others, on the contrary, are conserved because of the structural role they play in the molecule. Such is the case of the two prolines in this particular sequence: Pro40 produces a bending in a long alpha helix, while Pro76 is found in a particular conformation, favoring the establishment of the alpha helix that follows. The conservation of glycines is due to their lack of side chain, which allows them to adopt extreme conformations of their backbone angles (for example, Gly84 and Gly92), or to be accommodated in a very reduced space (Gly33). In another example, a polar interaction between Asp26 and Lys57 explains the amino acid profile observed at these positions. All this information we obtain by analyzing the structure testifies to the intimate relationship between structural role and evolutionary conservation.

From the early substitution models that considered evolution of sequences as the result of a homogeneous, purely neutral evolutionary process, models of molecular evolution have been improved substantially by including a variety of biological phenomena. Practically all the assumptions made for the Jukes-Cantor model (1969) have been relaxed in subsequent works, for the three types of sequence data (nucleotide, codon and amino acid). For example, heterogeneity in the substitution process, both across sites (see below) and over time (Yang and Roberts, 1995; Galtier and Gouy, 1998; Galtier, 2001; Huelsenbeck, 2002; Foster, 2004; Blanquart and Lartillot, 2006; Boussau and Gouy, 2006; Gowri-Shankar and Rattray, 2007; Zhou et al., 2010), has been introduced

in a number of models, producing almost invariantly an improvement in terms of fit.

Once again, we are interested in selective constraints pertaining to the protein structure, so I will focus on model extensions to address them, and omit details on the rest. As shown in figure 1.1, tertiary structure induces heterogeneity in the substitution process across sites. The first significant improvement in evolutionary models implicitly addressing this issue was the introduction of the *rates across sites* models (Olsen, 1987; Yang, 1993, 1994, 1996), where the rate of evolution is represented by a random variable drawn from a gamma distribution, or a discrete version of this distribution with a limited number of classes. In a similar perspective, the ω parameter of codon models can be drawn from various probability distributions to describe among-site variation of selective pressure (reviewed in Anisimova and Kosiol, 2009). In the simplest versions, a prespecified number of site classes is used (typically 3: positive selection, neutral and negative selection). Discrete versions of continuous distributions or distribution mixtures have also been applied (Yang et al., 2000; Pond and Muse, 2005). Yang and Swanson (2002) implemented models for prepartitioned datasets, for the case where prior information is available to partition sites in the protein into different classes. In another approach, Huelsenbeck et al. (2006) proposed the use of a Dirichlet process prior to model site-specific variation of ω; under this process, the number of classes is not predetermined, but it is instead a random variable controlled by a parameter estimated from the data.

In all the modeling alternatives described so far, however, only the overall rate of substitution but not the remaining parameters of the evolutionary model (equilibrium frequencies and relative rates of substitution) are allowed to vary across sites. We know, however, as illustrated in figure 1.1, that the heterogeneity in the substitution process is evident not only in the number of nonsynonymous substitutions, but also in the nature of these substitutions. This is the case for the stabilizing polar interaction between Asp26 and Lys57, which induces a particular amino acid profile for those positions in the alignment. Or for Leu99 and Leu103, whose importance for establishing the hydrophobic core of the molecule (Santos et al., 2007) prevents the fixation of non-hydrophobic

residues at these positions. A number of evolutionary models allowing consideration of changes in the substitution process (other than the rate parameter) have been proposed. In the model proposed by Bruno (Bruno, 1996; Halpern and Bruno, 1998), a vector of amino acid equilibrium frequencies specific for each site is considered. This approach requires a very large number of species in the alignment, since the amino acid frequencies have to be estimated for each column. The CAT model (Lartillot and Philippe, 2004), a mixture model allowing for a finite number of classes characterized by its own set of equilibrium frequencies, proved to be a more reasonable approach. In this model, a Dirichlet process prior is used to estimate the total number of classes and their respective amino acid profiles, with the class assigned to each site also a free parameter of the model. This approach has been recently implemented at the codon level (Rodrigue et al., 2010).

Explicit treatment of protein structure

The models described so far are phenomenological in nature: they capture substitution patterns through parameters estimated from the data, without an explicit modeling of the underlying causes (Rodrigue and Philippe, 2010). In the work presented here, on the contrary, we want to explore a mechanistic alternative, where prior knowledge on protein structure is directly incorporated, with the belief that the insights and advances of the structural biology community over the last years should help improving our understanding of sequence evolution.

Several attempts have been made to model evolution at the amino acid level with explicit treatment of structural constraints. Relaxing the assumption of a single exchangeability matrix for all sites, substitution matrices specific for predefined structural classes have been proposed (Overington et al., 1990; Wako and Blundell, 1994a,b; Koshi and Goldstein, 1995), with the implicit assumption that the structural environment of a residue is the main force acting on the evolution of this site. Dimmic et al. (2000) have extended this model, using a fitness function different for each one of a fixed number

of classes, where each site's class is a priori unknown. The relationship of the site classes with the protein structure is however not clear, because the many selective constraints operating at different sites are confounded. As a result, the correlation of the optimized fitness parameters of the model and the biophysical characteristics of amino acids is poor, and the interpretation of the fitness classes obtained is not evident. Thorne, Goldman and coworkers proposed probabilistic models where a Markov chain describes features of the secondary structure of proteins, and each category of structural environment uses a different Markov process model of amino acid replacement (Goldman et al., 1996; Thorne et al., 1996; Lio et al., 1998). The models provide an improvement in the description of the evolutionary process. However, only extremely simple structural representations have been used, namely a few categories of secondary structure and two states for solvent accessibility.

More importantly, all of the models described so far make the assumption of independence between sites, a simplification invoked for computational reasons but incompatible with a realistic treatment of the protein structure. It was not until recently that site interdependencies could be treated within a standard phylogenetic framework, which is the subject of the next section.

Structurally constrained evolutionary models

Modeling site dependencies in a probabilistic phylogenetic context is not a trivial task. Likelihood calculations using Felsenstein's pruning algorithm (Felsenstein, 1981) require the determination of transition probabilities between states, which involve rate matrix exponentiation (equation 1.6). When considering general site dependencies, the rate matrix is no longer a $4x4$, a $20x20$, nor even a $61x61$ matrix. If we assume that a substitution at one site may affect any other site in the molecule (which is not a very bold assumption from a biological perspective), the Markov process specified at the codon level is, in fact, equivalent to the process generated by a $61^N x 61^N$ matrix, with single entries describing rates of change from one N-codon sequence to another. The computational

cost of numerically calculating the transition probabilities with these high dimensional matrices is prohibitively expensive, and has justified the assumption of independence between sites usually invoked in phylogenetic methods.

Alternative techniques of likelihood calculations for evolutionary inferences dealing with dependence among changes at different positions have recently been proposed (Jensen and Pedersen, 2000; Hwang and Green, 2004; Siepel and Haussler, 2004; Christensen et al., 2005). Matrix exponentiation when calculating transition probabilities (equation 1.6) aims at integrating over all possible substitution histories, over a given phylogenetic phylogenetic tree. An alternative way of calculating this integral is to use MCMC to directly sample the complete substitution history, estimating in this way the value of this integral. Although still computationally demanding, the problem becomes now tractable.

These approaches were formulated to deal with context-dependent mutation, but Robinson et al. (2003) adapted the ideas of Jensen and Pedersen (2000) to the case where general dependencies are due to natural selection on phenotype, to explicitly model structural constraints within a standard phylogenetic framework. Their description of structural constraints is based on the work of Parisi and Echave (2001), who developed a technique for simulating the evolution of sequences that conform to a known tertiary structure. In this model, a scoring system for sequence-structure compatibility is used to evaluate the probability of fixation of a given mutation, assuming a coarse-grained protein structure that is constant through evolution. Nonsynonymous changes that make the sequence less compatible with the protein structure (for example, by introducing destabilizing interactions) will have a lower rate of occurrence. Formally, the instantaneous rate matrix originally proposed by Robinson et al. (2003) has the form

$$Q_{DEP_{ab}} = \begin{cases} 0 & \text{if } a \text{ and } b \text{ differ by more than one codon position} \\ \mu\pi_{c(b)} & \text{for a synonymous transversion} \\ \mu\pi_{c(b)}\kappa & \text{for a synonymous transition} \\ \mu\pi_{c(b)}\omega e^{\beta\Delta E(a,b)} & \text{for a nonsynonymous transversion} \\ \mu\pi_{c(b)}\kappa\omega e^{\beta\Delta E(a,b)} & \text{for a nonsynonymous transition} \end{cases} \tag{1.17}$$

Except for the term related to the protein structure in nonsynonymous rates, the parameterization is equivalent to the GY model described in equation 1.13. Parameters associated to the mutation process include the nucleotide equilibrium frequencies $(\pi_i)_{1\leq i\leq 4}$, the transition/transversion rate κ, and the parameter μ to scale the overall rate of change.

As for selective constraints acting on nonsynonymous changes, two terms are involved. The parameter ω, formulated in the same spirit as in the GY model, is intended to capture contributions to nonsynonymous rates that are not exclusively related to the protein structure. The innovation of this model lies in the term $e^{\beta\Delta E(a,b)}$, describing the effects of constraining the sequences to a particular protein structure. This term has two components. For a proposed sequence s, $E(s)$ measures how well s fits the protein structure. In their original formulation, Robinson et al. (2003) used a statistical potential originally derived for the protein-fold prediction problem (Jones et al., 1992a). The development of an accurate sequence-structure compatibility score for this type of evolutionary models is the main subject of this dissertation, so I will save the details for next sections. The second component of the structural term in this model, β, is treated as a free parameter, and estimated from the phylogenetic data along with the others. It represents the strength of selection for structural compatibility: when $\beta = 0$, the model simplifies to the widely used GY codon model (equation 1.13). Biologically reasonable

values of β are positive, corresponding to the case where evolution favors substitutions that fit the structure better. The higher the value of β, the stronger the role the structural term plays in the evolutionary model.

The focus of the works of Parisi and Echave (2001) and Robinson et al. (2003), and the subsequent work of Rodrigue and coworkers (Rodrigue et al., 2005, 2006, 2007) was on the definition of the evolutionary model and the statistical tools to perform phylogenetic inference and model comparison dealing with site dependencies. The sequence-structure compatibility measure, however, was to some extent neglected. In its most complex form, an empirical potential originally derived for the protein-fold prediction problem (Jones et al., 1992a) was used, consisting in two components: one accounting for solvent accessibility requirements, and the other related to pairwise interactions between residues close in space. Since, in contrast to the phenomenological models presented in previous sections, this approach attempts to provide a mechanistic description of the way natural selection operates on the evolutionary process, the accuracy of the model will be highly dependent on how this mechanistic description matches reality. In the following chapters, we will concentrate on ways of measuring how well a sequence fits a structure, and how this measure can be improved without incurring in excessive computational costs.

1.2 Evaluating sequence-structure compatibility

In the phylogenetic framework just described, each mutation undergone by a protein during evolution has to be evaluated for its compatibility with the structure and contrasted with all the other possible mutations, assuming that the tertiary structure remains invariant. This formulation presents important analogies to the protein design problem, where the goal is to find sequences that fold into a given conformation. Since the size of both the sequence and conformational space are extremely large, there is a trade-off between accuracy and speed when evaluating each sequence, and protein design ap-

proaches use several strategies to speed up the scoring process. This trade-off is even more pronounced in the phylogenetic context, because of the additional computational burden involved in calculating likelihood scores.

There are two very different types of scoring functions currently used. The first ones are physical energies that can be obtained, in principle, from a fundamental analysis of forces between particles. The second ones, called statistical potentials, work with simplified versions of the proteins, and their parameters are derived from known protein structures.

1.2.1 Physical energies

Semi-empirical potentials, widely used to perform molecular mechanics calculations, such as CHARMM (Brooks et al., 1983), AMBER (Cornell et al., 1995) and OPLS (Jorgensen and Tirado-Rives, 1988), work at the atomic level. They consist of a mathematical expression of the energy of a system as a function of the cartesian coordinates of the atoms ($\overrightarrow{R}$). Although quantum mechanical calculations can yield potential surfaces for small molecules, it is not yet feasible to calculate directly such surfaces for large macromolecules. Semi-empirical approaches work, instead, with a fairly simple, though atomically detailed, 'ball and spring' type models: atoms are represented as spheres with point charges, with chemical bonds treated as springs. In addition to the atomic coordinates, thus, the energy value also depends on a set of parameters that describe the geometric and energetic properties of interactions between particles, adjusted to optimize agreement with experimental data and with quantum calculations on smaller molecules (Karplus and Petsko, 1990; Ponder and Case, 2003; Guvench and MacKerell, 2008). The combination of the mathematical function and the parameters is commonly referred to as a "force field".

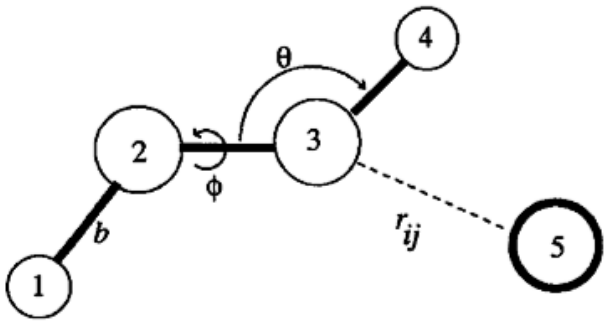


Figure 1.2: **Schematic view of force field interactions.** Covalent bonds are indicated by heavy solid lines, nonbonded interactions by a light, dashed line. Figure from Ponder and Case (2003).

The most commonly used protein force fields incorporate a relatively simple potential energy function (Ponder and Case, 2003):

$$V(\overrightarrow{R}) = \sum_{bonds} k_b(b-b_0) + \sum_{angles} k_\theta(\theta-\theta_0) + \sum_{torsions} k_\phi[cos(n\phi+\delta)+1] + \sum_{nonbond\,pairs} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{Cij}{r_{ij}^{6}}\right] \tag{1.18}$$

The first three summations are over bonds (interactions between atoms 1-2 in figure 1.2, angles (interactions between atoms 1-3), and torsions (between atoms 1-4). The final sum, over all pairs of atoms i and j, describes electrostatics that use partial charges q_i on each atom that interacts via Coulomb's law. The combination of dispersion and exchange repulsion forces are represented by a Lennard-Jones 6-12 potential; this is often called the 'van der Waals' term. Equation 1.18 is about the simplest potential energy function that can reproduce the basic features of protein energy landscapes at an atomic level of detail (Ponder and Case, 2003).

When calculating the energy of a protein upon mutation, two main requirements of such detailed molecular representation involve high computational costs. First, the position of every atom of the new sequence has to be determined. Even when reducing the conformational search by assuming a fixed backbone, we still have to face the prob-

lem of positioning side-chains. The complexity of this search can be further reduced through the use of rotamer libraries (collections of statistically preferred side chain conformations for each residue type (Tuffery et al., 1991; Dunbrack and Karplus, 1993), but it still implies large computational times, and the energy function has to be adjusted *ad hoc* to accommodate such simplifications. Second, there is a need for an accurate treatment of the solvent environment, which adds a significant number of atoms to the system. Such treatment may be performed using explicit (that is, modeling separately each solvent molecule) or implicit models (Roux and Simonson, 1999), with the former being a more microscopically complete method while the latter having the advantage of savings in computer time. For the CHARMM potential, for example, the calculations with explicit water molecules are approximately 200-500+ times slower than the corresponding vacuum calculations. The implicit solvent models, in turn, imply a reduction of speed of 1.5 to 175 times with respect to vacuum (Brooks et al., 2009).

Computational costs aside, the application of these semi-empirical potentials to answer evolutionary questions warrants further considerations. While they provide a sound theoretical basis for calculating energy changes after a substitution, they require a precise definition of the system (protein, solvent and conditions), and are sensitive to the simplifying assumptions needed to make a sequence search problem tractable (Gordon et al., 1999). Their underlying hypothesis is that the behavior of proteins can be described in terms of the basic physical principles governing their elementary atomic constituents (Brooks et al., 2009). Accordingly, parameterizations of these potentials use fits to quantum calculations or empirical data on very simple systems, sometimes developed and tested primarily on gas-phase simulations. While empirical potentials for gas-phase, non-polar organic molecules are extremely accurate, and a molecular mechanics computation is as trustworthy as the corresponding experimental results, the situation is currently much less satisfactory for proteins and complex systems (Ponder and Case, 2003). Even with the use of similar functional forms, different versions of the traditional force fields still exhibit significant differences in the results, as shown in simulations of

dipeptides in solution (Ponder and Case, 2003; Hu et al., 2003), and of a large set of mutations on complete proteins (Potapov et al., 2009). Evolutionary studies imply much more complex and poorly defined systems: the estimated energies should be valid in the context of a living cell, with crowded and changing environments. It is far from clear that the accuracy of traditional force fields would carry over to such systems.

1.2.2 Statistical potentials

An alternative to the semi-empirical strategy consists in the use of knowledge-based, or statistical potentials, which are derived from the analysis of known protein structures. The probabilities that residues appear in specific configurations (such as in buried or surface environments, or rotamer conformations) or that pairs of residues are found close in space are calculated. Knowledge based potentials are, thus, scoring functions that encode statistical patterns present in solved protein structures. They are inductive in nature, based on the idea that the propensity of an amino acid in a given site of a protein can be predicted by the observed frequency of that amino acid in other similar structural contexts in other proteins. They should in principle capture all kinds of patterns that biological sequences have, in relation to their conformation, and not only those directly related to thermodynamic stability. In spite of a lack of theoretical basis (Thomas and Dill, 1996b; Ben-Naim, 1997), statistical potentials implicitly account for complex effects, even when a good physical understanding of the underlying causes does not exist (Lazaridis and Karplus, 2000; Boas and Harbury, 2007).

Knowledge based potentials are extremely fast to compute. They are not restricted to all-atoms representations but can work instead with coarse grained versions of the structure, with an arbitrary level of detail in the description. The number of energetic calculations required is thus reduced: fewer points are considered, conformational space is discrete and restricted, and the total number of interactions is reduced -instead of considering hydrogen-bonding, van der Waals forces, etc., between the multiple atoms of amino acid residues, there is only a single energetic term for each possible residue

pairing. Individual amino acids are treated as single points on a chain, avoiding the problem of side-chain positioning.

This coarse grained representation, which provides a very low resolution of the protein structure and may be insufficient for many applications, offers several advantages in our context. When compared to atom-based physical energies, residue-based statistical potentials tend to present a smoother energy landscape, which makes them less sensitive to small displacements (Lazaridis and Karplus, 2000). As a result, they are more robust for low-resolution structure assessment, for which small errors are inevitable. In particular, in our evolutionary model, small errors are necessarily introduced by the assumption of a protein structure that remains invariant through evolution. Furthermore, using a coarse grained representation should avoid the problem of only accepting near-native sequences, i.e., sequences too close to the one corresponding to the reference structure. This is an effect observed when using an atomic representation combined with the assumption of a fixed backbone (Kuhlman and Baker, 2000), and may introduce artifacts in our context. Finally, the level of structural detail and the particular elements of the protein structure considered in the evolutionary model can be arbitrarily defined, providing a flexibility in the approach very difficult to obtain otherwise.

Statistical potentials are now widely used tools for several applications, such as assessment of experimetally determined and theoretically predicted protein structures (Melo et al., 2002), fold recognition or threading (Jones et al., 1992a), detection of native-like protein conformations (Gatchell et al., 2000), and protein design (Poole and Ranganathan, 2006). In contrast to semi-empirical physical energies, which have become fairly standardized, knowledge based potentials are extremely diverse. A great variety of potentials have been derived since the initial formulations (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985; Sippl, 1990), differing in the theoretical basis of their formulation, the optimization methods and the definition of interacting centers and types of interactions (reviewed in Lazaridis and Karplus, 2000; Poole and Ranganathan, 2006; Boas and Harbury, 2007; Rykunov and Fiser, 2010). The various terms

are typically calibrated and weighted to optimize performance in the specific application they were developed for, but there is a lack of reliable objective criterion for selecting an appropriate potential (Mirny and Shakhnovich, 1996).

In this context, we were interested in developing a statistical potential conceived with the phylogenetic inference framework in mind: accounting for properties of the protein structure that may be important at an evolutionary scale, while keeping the calculations tractable. In the next chapter of this dissertation, I present a probabilistic formulation for optimizing the parameters of a coarse-grained, residue level statistical potential. The method provides objective ways of selecting models for otherwise arbitrary definitions of the terms, with a formulation general enough to incorporate many of the potentials proposed over the last years. In the following chapter, this model selection framework is used to increase the level of detail of the structural representation, while keeping computation costs low. Always at the residue-level, several structural elements are investigated. The potentials are then included into a structurally constrained evolutionary model. Their performance is evaluated in terms of model fit, and contrasted against standard evolutionary models. Finally, this new structurally constrained phylogenetic model is used to understand the selective forces behind the differences in conservation found in genes of very different expression levels.

CHAPTER 2

A MAXIMUM LIKELIHOOD FRAMEWORK FOR PROTEIN DESIGN

In this chapter a general statistical framework for optimizing knowledge-based potentials conceived for evolutionary studies is presented. In evolution, a protein's structure changes very slowly, while a multitude of sequences produced by random mutations have to conform to this structure. We thus posed the problem in terms of protein design, or the inverse folding problem, that is, predicting sequences compatible with a given structure. A probabilistic formulation is developed, where the goal is to obtain a probability distribution of sequences conditional on a structure.

The main contribution of this work is to propose an overall statistical framework for the protein design problem, based on the Maximum Likelihood principle. This framework entails the following two main advantages over previous works: theoretical guarantees of learning optimality, and well defined methods for comparing the performances of alternative potentials of different forms.

Parameters of the potentials are optimized by maximizing the joint probability of observing the set of training proteins. The maximization is performed by gradient descent, with an MCMC procedure embedded to numerically estimate the derivatives of the log-likelihood function. In Appendix 1, an alternative formulation is presented that avoids the use of MCMC, thus markedly reducing computation times.

The goal of this article was to introduce the general methodology. The functional form of the potential chosen -a contact potential supplemented with solvent accessibility- is meant as an illustration, and is thus (maybe too) simple. More complex forms will be studied in chapter 3.

BMC Bioinformatics 2006, **7**:326; doi:10.1186/1471-2105-7-326

A maximum likelihood framework for protein design

Claudia L. Kleinman[1], Nicolas Rodrigue[1], Cécile Bonnard[2], Hervé Philippe[1] and Nicolas Lartillot[2]

1. *Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Québec Canada*
2. *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506,*
CNRS-Université de Montpellier 2, 161, rue Ada, 34392 Montpellier Cedex 5, France

ABSTRACT

Background: The aim of protein design is to predict amino-acid sequences compatible with a given target structure. Traditionally envisioned as a purely thermodynamic question, this problem can also be understood in a wider context, where additional constraints are captured by learning the sequence patterns displayed by natural proteins of known conformation. In this latter perspective, however, we still need a theoretical formalization of the question, leading to general and efficient learning methods, and allowing for the selection of fast and accurate objective functions quantifying sequence/structure compatibility.

Results: We propose a formulation of the protein design problem in terms of model-based statistical inference. Our framework uses the maximum likelihood principle to optimize the unknown parameters of a statistical potential, which we call an *inverse potential* to contrast with classical potentials used for structure prediction. We propose an implementation based on Markov chain Monte Carlo, in which the likelihood is maximized by gradient descent and is numerically estimated by thermodynamic integration. The fit of the models is evaluated by cross-validation. We apply this to a simple pairwise contact potential, supplemented with a solvent-accessibility term, and show that the resulting models have a better predictive power than currently available pairwise potentials. Furthermore, the model comparison method presented here allows one to measure the relative contribution of each component of the potential, and to choose the optimal number of accessibility classes, which turns out to be much higher than classically considered.

Conclusions: Altogether, this reformulation makes it possible to test a wide diversity of models, using different forms of potentials, or accounting for other factors than just the constraint of thermodynamic stability. Ultimately, such model-based statistical analyses may help to understand the forces shaping protein sequences, and driving their evolution.

2.1 Background

Predicting the sequences compatible with a given structure defines what is traditionally called the inverse folding problem, or more often, protein design (Drexler, 1981; Pabo, 1983; Ponders and Richards, 1987). As suggested by the terminology, this question is usually considered in an engineering perspective: the aim is then to determine a sequence, or a set of sequences, that stably fold into a pre-specified conformation. In a thermodynamic perspective, this requirement translates into eliciting sequences that have lowest free energy under the target fold, compared to all possible alternative conformations. In principle, such a criterion would imply a search through the joint structure-sequence space, which is not feasible but for small on-lattice model proteins (Seno et al., 1996).

As an alternative to the engineering approach, a more evolutionary stance can be taken towards the inverse folding problem, in which case the aim would rather be to predict the sequences of *natural* proteins having the conformation of interest. Seen from this new point of view, the design problem raises new questions: natural proteins are the result of a complex evolutionary process, involving an intricate interplay between mutation and selection, and this probably entails many constraints directly related to the native conformation, but nevertheless not equivalent to the mere requirement of structural stability. For instance, the requirement of fast and cooperative folding has an impact on the dispersion of contact energies (Abkevich et al., 1996). For this and many other potential reasons, among all sequences predicted by classical engineering-oriented protein design, probably only a subset will look like natural proteins.

The evolutionary approach to protein design is particularly relevant to phylogenetic studies, where one of the current motivations is to develop the so-called structurally constrained models of protein evolution, i.e. models explicitly dependent on the protein's conformation, either for simulation purposes (Hellinga and Richards, 1994; Parisi and Echave, 2001; Bastolla et al., 2002, 2003), or in the context of phylogenetic infer-

ence (Robinson et al., 2003; Rodrigue et al., 2005). In this framework, each substitution undergone by a protein during evolution has to be tested for its compatibility with the structure, in the context of the sequence that the protein displays at all other sites when the substitution occurs. Such repeated evaluation of the structure-sequence compatibility along a phylogenetic tree requires relevant and computationally very efficient scoring schemes/functions.

It is interesting to compare the different methods proposed thus far for performing protein design in light of this engineering/evolutionary distinction. A first direction of research has consisted in using all-atom semi-empirical force fields to evaluate the conformational free energy (reviewed in Park et al., 2004). These empirical methods have been applied to many theoretical and experimental cases, reaching a high level of accuracy. On the other hand, they are computationally heavy, mainly because of the side-chain positioning problem, and thus cannot be easily applied to structurally constrained phylogenetic models (Robinson et al., 2003; Rodrigue et al., 2005). Concerns may also be expressed about their over-sensitivity to the native conformation, in particular in the core of the target structures and when the flexibility of the backbone is not accounted for (Wernisch et al., 2000; Larson et al., 2002). But more importantly, approaches based on physical force fields are, by definition, exclusively focussed on the conformational stability, and thereby, completely oversee other potential factors shaping the sequences of biological proteins. As such, they are well suited for engineering synthetic proteins (Dahiyat et al., 1997), or for testing to what extent natural sequences are shaped by selection for protein stability (Jaramillo et al., 2002), but may not be sufficient for more general evolutionary purposes.

An alternative to the semi-empirical strategy consists in relying on knowledge-based, or statistical, potentials. These scoring functions mimic physical Boltzmann distributions, but merely encode statistical patterns present in the databases. Some of these potentials were obtained under the quasi-chemical approximation, whereby frequencies of patterns, such as contacts between each pair of amino-acids, are transformed into

energies using the Boltzmann law (Miyazawa and Jernigan, 1985; Sippl, 1993; Godzik et al., 1995; Solis and Rackovsy, 2006). Alternatively, contact energies can be obtained by maximizing the potential's predictive accuracy in a threading test (Hendlich et al., 1990; Maiorov and Crippen, 1992; Mirny and Shakhnovich, 1996; Bastolla et al., 2001). In the present context, an advantage of these knowledge-based potentials, compared to semi-empirical force-fields, is that they should in principle capture all kinds of patterns that true biological sequences have, in relation to their conformation, and not only those directly related to thermodynamic stability. Furthermore, statistical potentials need not be defined at the atomic level, but can be based on a coarse-grained description of the protein's configuration, essentially by omitting the degrees of freedom associated to side chains. This allows faster computations, by avoiding the problem of searching through the rugged landscape of side-chain conformations. In addition, coarse-grained potentials could turn out to be an advantage, in that they will not recover the native sequence too faithfully. Most protein design procedures based on statistical potentials proposed until now have relied on coarse-grained, pairwise contact pseudo-energies (Shakhnovich and Gutin, 1993; Kurosky and Deutsch, 1995; Deutsch and Kurosky, 1996; Seno et al., 1996, 1998; Micheletti et al., 1998; Banavar et al., 1998; Rossi et al., 2000, 2001).

Yet, irrespective of the level of description adopted, currently available statistical potentials may not be ideal for protein design, since they have generally been optimized in the context of the folding problem, i.e. for maximizing the rate of correct structure prediction, given the sequence. In contrast, we would like to optimize the reciprocal prediction, namely, the sequences given the conformation. Several approaches have been proposed in this direction, consisting in maximizing the Z-score between the energy of the native sequence on the target conformation and its energy on a set of decoy sequences (Chiu and Goldstein, 1998), or, alternatively, in applying a mean-square criterion on the values taken by the scoring function on each structure-sequence pair of the database (Seno et al., 1998). However, these methods have thus far only been tested in cubic lattice protein models. In addition, they lack a firm theoretical basis. In particu-

lar, it would be interesting to guarantee optimal predictive power, and to have a robust methodology available to assess and compare the performance of alternative forms of statistical potentials.

Standard statistical theory provides such theoretical guarantees (Wald, 1949). In the present case, the inverse folding problem can be formulated directly in terms of the probability of observing a sequence s given a conformation c, i.e. $p(s \mid c, \theta)$. This probability explicitly depends on the pre-specified model through a series of parameters, represented here by θ. These may be, for instance, the coefficients of a pairwise potential, parameters describing compositional effects, secondary structure environment, solvent accessibility, etc. Taking the product over a database of P independent sequence-conformation pairs, $S = (s^p)_{p=1..P}$ and $C = (c^p)_{p=1..P}$, yields a joint probability

$$p(S \mid C, \theta) \quad = \quad \prod_p p(s^p \mid c^p, \theta) \tag{2.1}$$

which, as a function of θ, can be seen as a likelihood. The parameter θ is then learnt by maximizing the likelihood with respect to θ. Once this is done, sequences can be assessed, or sampled, under the optimal parameter value $\hat{\theta}$, by direct numerical evaluation of their probability, or by Monte Carlo sampling methods.

Reformulated in this way, the method maximizes the predictive power of the potential, now in the structure-seeks-sequence direction. By construction, it yields the optimal parameter values that can be obtained for a given form of the potential. In addition, the fit of the model can be directly evaluated, based on the value of the likelihood obtained on a test data set, distinct from the learning set (cross-validation), giving a means of rigorous model selection. Finally, the statistical framework proposed here allows one to explicitly combine together, in a model dependent manner, all kinds of factors that we surmise may induce correlations between the structure and the sequence of proteins.

We have implemented this maximum likelihood (ML) procedure in a Markov chain Monte Carlo framework, and applied it to a simple case, using a contact potential, sup-

plemented with a solvent accessibility term. Using cross-validation, we show that the resulting potentials yield a better fit than currently available potentials of the same form, and that combining solvent-accessibility considerations with contact energies is better than either alone. Furthermore, we find that solvent accessibility requires a more complex description than what is currently used. Ultimately, the overall method proposed in this work can be extended to a large spectrum of alternative models and statistical potentials.

2.2 Results

2.2.1 The probabilistic model

Let us consider a sequence $s = (s_i)_{i=1..N}$, of length N, and of conformation c. In its most general form, the method introduced here can work with any model M specifying the conditional probability of s given c, in terms of an unnormalized non negative function $q(s,c)$:

$$p(s \mid c, M) = \frac{q(s,c)}{\sum_s q(s,c)}. \tag{2.2}$$

To illustrate the method, we will apply it to a simple case, using a pairwise contact potential. The argument is as follows. First, by Bayes' theorem:

$$p(s \mid c, M) = \frac{p(c \mid s, M)\ p(s \mid M)}{\sum_s p(c \mid s, M)\ p(s \mid M)}. \tag{2.3}$$

If, in addition, we assume a uniform prior on s, we can simply relate equations 2.3 and 2.2 by posing $q(s,c) = p(c \mid s, M)$. Next, given a statistical potential $E(s,c)$, the conformational probability $p(c \mid s)$ can be expressed as a Boltzmann distribution:

$$p(c \mid s, M) = \frac{e^{-E(s,c)/kT}}{Z_s} \tag{2.4}$$

$$= e^{-(E(s,c)-F(s))/kT}, \tag{2.5}$$

where

$$Z_s = \sum_c e^{-E(s,c)/kT} \tag{2.6}$$

is a normalization constant, and

$$F(s) = -\ln Z_s. \tag{2.7}$$

T and k are the absolute temperature and the Boltzmann constant, respectively. Without loss of generality, it is possible to rescale the potential so that $kT = 1$, which we will do in the following. Then, by defining the *inverse potential*:

$$G(s,c) = E(s,c) - F(s), \tag{2.8}$$

the conditional probability of sequence s reads as

$$p(s \mid c, \theta, M) = \frac{e^{-G(s,c)}}{Y}, \tag{2.9}$$

where

$$Y = \sum_{s'} e^{-G(s',c)} \tag{2.10}$$

is the normalization factor. Note that, contrary to the Z_s factor of equation 2.4, which was a sum over all conformations, the present factor Y is a sum over sequence space (all possible sequences of length N).

2.2.2 Statistical potentials

In the present work, we used a statistical potential made of two terms:

$$E(s,c) \quad = \quad \sum_{1\leq i<j\leq N} \Delta_{ij}\varepsilon_{s_i s_j} + \sum_{1\leq i\leq N} \alpha_{s_i}^{v_i}. \tag{2.11}$$

The first term is a contact free energy: $\Delta_{ij} = 1$ if positions i and j are closer in space than a certain cut-off distance, and 0 otherwise, and ε_{ab} defines the contact energy between amino acids a and b. The second term encodes a solvent-accessibility free energy: for each position, α_a^d represents the free energy of amino acid a in the solvent accessibility class d, $a = 1..20$, and $d = 1..D$, where D is the total number of solvent accessibility classes considered.

Deriving the inverse potential requires the calculation of $F(s)$, which is already entirely specified by the potential E as a sum over all conformations. However, this computation is difficult in practice. As an alternative, we can give it a simple phenomenological form, inspired from the random energy model (Shakhnovich and Gutin, 1993; Sun et al., 1995; Seno et al., 1998):

$$F(s) \quad = \quad -\sum_{1\leq i\leq N} \mu_{s_i}, \tag{2.12}$$

where the $(\mu_a)_{a=1..20}$ are unknown parameters, analogous to "chemical potentials" for the 20 amino acids.

Altogether, our parameter vector is made of three components: $\theta = (\alpha, \varepsilon, \mu)$, and the inverse potential reads as:

$$G(s,c) \quad = \quad \sum_{1\leq i<j\leq N} \Delta_{ij}\varepsilon_{s_i s_j} + \sum_{1\leq i\leq N} \alpha_{s_i}^{v_i} + \sum_{1\leq i\leq N} \mu_{s_i}. \tag{2.13}$$

Note that the probability defined by equation 2.9 is invariant under the following

transformation:

$$\mu'_a = \mu_a + J_1, \tag{2.14}$$

$$\varepsilon'_{ab} = \varepsilon_{ab} + J_2, \tag{2.15}$$

$$\alpha'^{d}_{a} = \alpha^d_a + J_3, \tag{2.16}$$

where J_1, J_2 and J_3 are arbitrary real constants. Therefore, to ensure identifiability of our probabilistic model, we enforce the following constraints:

$$\sum_a \mu_a = 0, \tag{2.17}$$

$$\sum_{ab} \varepsilon_{ab} = 0, \tag{2.18}$$

$$\sum_a \alpha^d_a = 0\,, d = 1..D. \tag{2.19}$$

A series of alternative inverse potentials can be obtained by suppressing the first or the second of the components of equation 2.13. In the present work, we tested the following combinations: μ, $\alpha+\mu$, $\varepsilon+\mu$, $\varepsilon+\alpha+\mu$.

We also explored various numbers of accessibility classes, with D ranging from 2 to 20. Alternatively, the ε component can be fixed to values of a contact potential obtained by other authors (MJ) (Miyazawa and Jernigan, 1985). In this case, we must add a multiplicative scaling factor λ in front of the contact component to account for the fact that these potentials are normalized differently:

$$G(s,c) = \lambda \sum_{1\le i<j\le N} \Delta_{ij}\varepsilon^{MJ}_{s_i s_j} + \sum_{1\le i\le N} \mu_{s_i}. \tag{2.20}$$

The scaling factor is optimized by ML, along with μ.

2.2.3 Optimizing the potentials by gradient descent

If we now consider a database, made of P protein sequences $S = (s^p)_{p=1..P}$, of respective lengths N_p and their corresponding three dimensional structures $C = (c^p)_{p=1..P}$, the probability of observing the whole database, which we define as the *likelihood* $L(\theta)$, is the product of the probabilities of observing each protein independently:

$$\begin{aligned} L(\theta) &= p(S \mid C, \theta) && (2.21)\\ &= \prod_p p(s^p \mid c^p, \theta) && (2.22)\\ &= \frac{e^{-G(S,C)}}{Y} && (2.23) \end{aligned}$$

where

$$G(S,C) = \sum_p G(s^p, c^p) \quad (2.24)$$

is the inverse potential summed over the database, and

$$Y = \sum_{S'} e^{-G(S',C)} \quad (2.25)$$

is the corresponding normalization constant. Since it is more convenient to work on minus the logarithm of the probability, we define the score ω:

$$\begin{aligned} \omega(\theta) &= -\ln L(\theta) && (2.26)\\ &= G(S,C) + \ln Y. && (2.27) \end{aligned}$$

We wish to maximize the likelihood, or equivalently, minimize ω, with respect to θ. We do this by gradient descent, based on a numerical evaluation of the derivative of ω (see methods). The overall method is akin to an Expectation Maximization algorithm (Dempster et al., 1977). In fact, it can be seen as a differential version of Demp-

ster's method, and therefore, we call it *differential EM*.

The derivative of ω reads as:

$$\frac{\partial \omega}{\partial \theta} = \frac{\partial G(S,C)}{\partial \theta} + \frac{\partial \ln Y}{\partial \theta}. \tag{2.28}$$

Applying the partition function formalism to equation 2.25, we can express the second term as an expectation over $p(S' \mid C, \theta)$:

$$\begin{aligned}
\frac{\partial \ln Y}{\partial \theta} &= \frac{1}{Y}\frac{\partial Y}{\partial \theta} && (2.29)\\
&= -\frac{1}{Y}\sum_{S'} \frac{\partial G(S',C)}{\partial \theta} e^{-G(S',C)} && (2.30)\\
&= -\sum_{S'} \frac{\partial G(S',C)}{\partial \theta} p(S' \mid C, \theta) && (2.31)\\
&= -\langle \frac{\partial G}{\partial \theta} \rangle && (2.32)
\end{aligned}$$

which leads us to the following expression for the derivative of ω:

$$\frac{\partial \omega}{\partial \theta} = \frac{\partial G(S,C)}{\partial \theta} - \langle \frac{\partial G}{\partial \theta} \rangle. \tag{2.33}$$

The computation of the first term in this equation is straightforward, while the second term must be estimated numerically. In order to do so, we obtain a sample $(S_h)_{h=1..K_{EM}}$ drawn from $p(S \mid C, \theta)$ by a Gibbs sampling algorithm similar to that of Robinson et al. (2003) (see methods).

Applying formula 2.33 on the inverse potential 2.13 yields the following expressions for the derivatives:

$$\frac{\partial \omega}{\partial \varepsilon_{ab}} = -\left[n_{ab} - \langle n_{ab} \rangle\right], \tag{2.34}$$

where n_{ab} is the number of contacts between amino acids a and b observed in the

database, and $\langle n_{ab} \rangle$ is its expectation over the probability distribution $p(S' \mid C, \theta)$. Formula 2.34 thus leads to an intuitive characterization of the maximum likelihood estimate $\hat{\varepsilon}$: it is the value of ε such that the average number of each type of contact predicted by the potential matches the number observed in the database. Following a similar derivation:

$$\frac{\partial \omega}{\partial \mu_a} = -\left[m_a - \langle m_a \rangle\right], \tag{2.35}$$

where m_a is the total number of amino acids of type a, and

$$\frac{\partial \omega}{\partial \alpha_a^d} = -\left[l_a^d - \langle l_a^d \rangle\right], \tag{2.36}$$

where l_a^d is the total number of amino acids of type a belonging to solvent-accessibility class d.

We first performed an optimization of the pure contact potential ($\varepsilon + \mu$-potential) on each data set. Figure 2.1 shows the evolution of the scoring function ω and of the contact potential during the gradient descent. As can be seen from these traceplots, the differential EM algorithm converges after a few hundred cycles. The scoring function stabilizes at around 272,000 natural units of logarithm (nits), and then fluctuates by up to 25 nits around this value. These fluctuations are mainly due to the finite size of the sample of sequences on which the derivative of $\ln Y$ is evaluated and, to a lesser extent, to the error on the estimation of $\ln Y$ by thermodynamic integration. In any case, these errors are small compared to the differences between scores obtained with alternative models (see below).

The evolution of the potential for some residue pairs is shown in figure 2.1b and 2.1c. Effects in the final values due to residue polarity are easily seen: known favorable interactions such as glutamate-lysine or the hydrophobic isoleucine-valine have a lower contact energy, while known unfavorable interactions, such as glutamate-glutamate, have

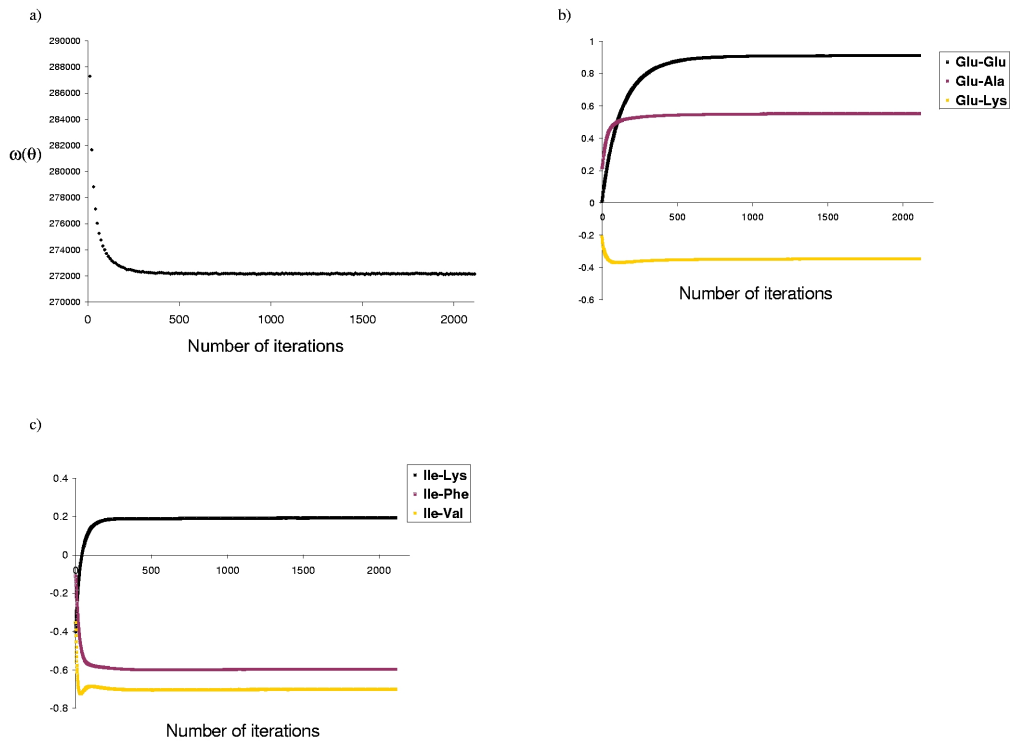


Figure 2.1: **Convergence of the optimization procedure.** (a)Traceplots illustrating the convergence of the differential EM method in the optimization of contact potentials, on data set DS1. Are shown, as a function of the number of iterations **(a)** the score $\omega(\theta) = -\ln p(S \mid C, \theta)$, **(b)** and **(c)** examples of pairwise contact energies obtained for some amino acid pairs.

higher energies, indicating that the potentials obtained are biologically reasonable.

The potentials obtained in two independent runs are virtually identical (figure 2.2a), indicating that the gradient descent does not get trapped into local minima. We can also compare the values of the potential for two distinct data sets of equivalent size, DS1 and DS2 (figure 2.2b), which uncovers a greater discrepancy than for two independent runs on the same data set DS1. The correlation is high, however, suggesting that data sets are large enough for the learning procedure to reach stability. In addition, these differences are small compared to the discrepancy between the potential obtained by our method and that of Miyazawa & Jernigan (figure 2.2c).

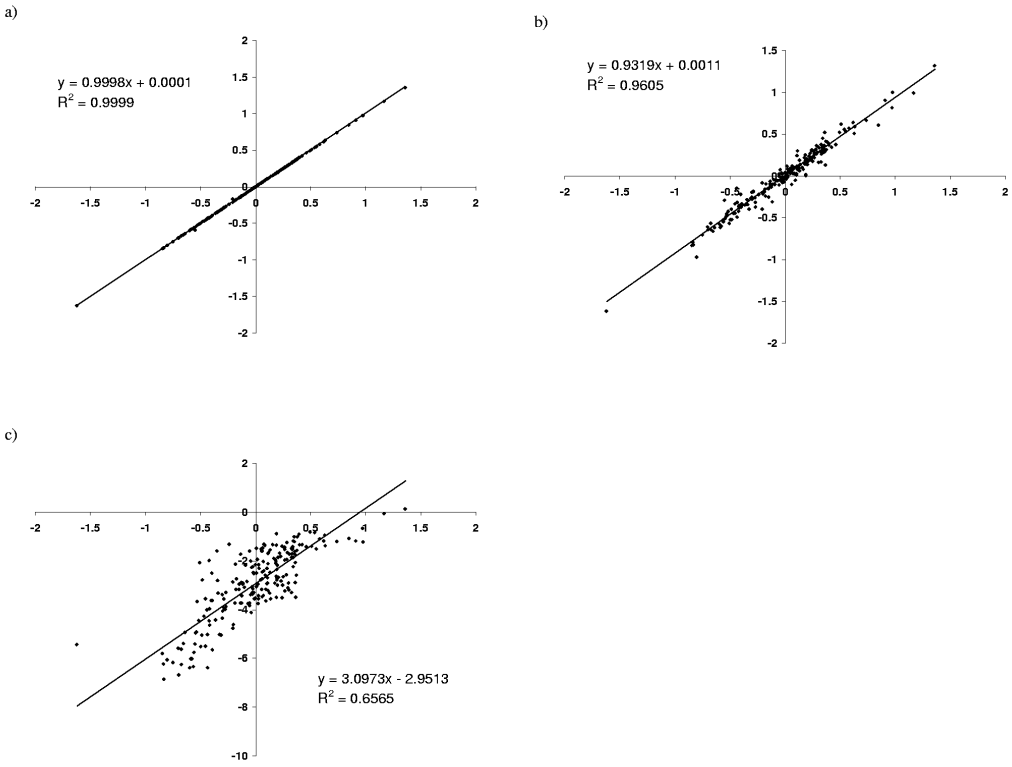


Figure 2.2: **XY-comparisons of pairwise contact potentials. (a)** two independent runs on the same data set DS1, **(b)** two runs, on data sets DS1 (X-axis) and DS2 (Y-axis); **(c)** Miyazawa and Jernigan's potential, compared to that obtained on DS1.

2.2.4 Model comparison

The same optimization procedure was applied to the potential consisting only of the solvent accessibility term ($\alpha + \mu$), with an increasing number of accessibility classes, and to the combined ($\varepsilon + \alpha + \mu$) potential. The resulting log likelihood scores cannot directly be compared, since the models do not have the same dimensionality. We therefore applied a 2-fold cross-validation procedure (CV), consisting in learning the potential on DS2, and testing it on DS1, and vice versa.

The evolution of the CV score as a function of the number of accessibility classes (D) is shown in figure 2.3. When D increases, the fit of the model improves, until

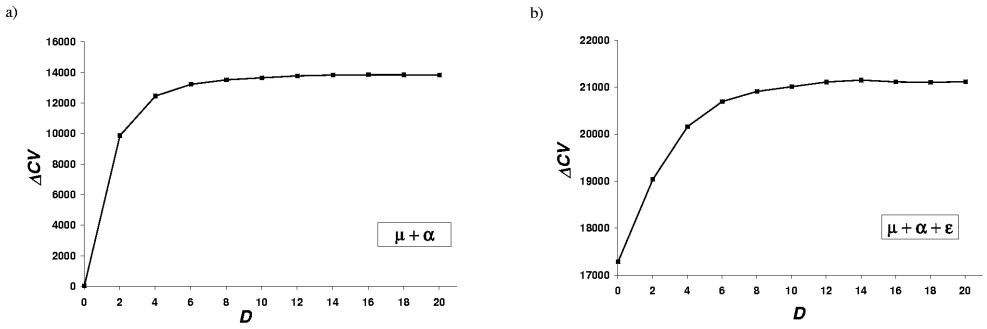


Figure 2.3: **Effect of the solvent accessibility definition on the potential.** Gain in cross-validation score (see Methods) as a function of the number of accessibility classes. The average gain for the 2-fold cross-validation experiment is shown. **(a)** Inverse potential consisting in solvent accessibility terms only, and **(b)** inverse potential combining contact and solvent accessibility terms.

a point is reached where the penalization for model dimensionality starts to dominate the score. The optimal number of classes obtained is 14 to 16, depending on the form of the potential studied, although 4 to 6 classes is sufficient to attain 90% of the fit improvement.

The scores obtained for the different models tested are reported in figure 2.4. We also included in the comparison the Miyazawa and Jernigan potential (Miyazawa and Jernigan, 1985). The contact potential performs better than the pure solvent accessibility potential, and the combination of both terms is the most informative. Miyazawa and Jernigan's potential results in a poorer fit improvement than any of the other models.

2.2.5 Specificity of the designed sequences

Once an optimal value of θ is obtained, properties of the sequences induced by the models can be investigated by sampling sequences from $p(s \mid c, \theta)$, using this optimal value of θ. In particular, we tested to what extent the sequences proposed by our method met the requirement of specificity, i.e. the condition that the sequences designed on a

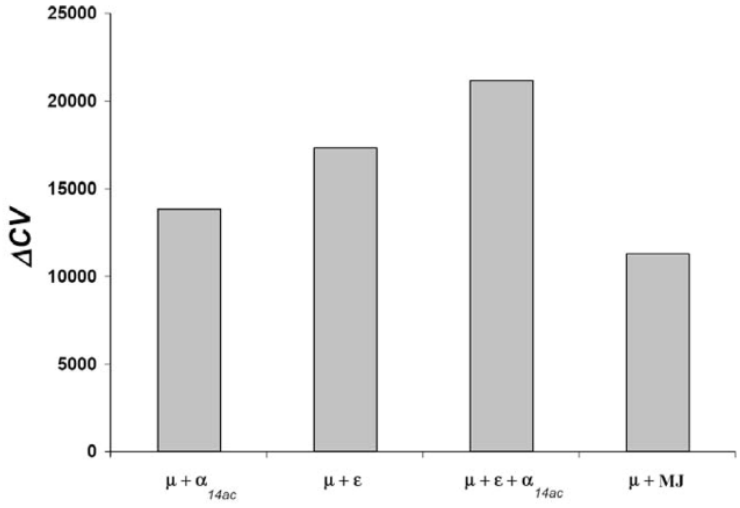


Figure 2.4: **Model comparison.** Cross-validation (CV) scores obtained for the different forms of potentials tested. The average gain (relative to the CV score obtained with the flat potential μ, see Methods) for the 2-fold cross-validation experiment is reported. α_{14ac}: solvent accessibility potential, 14 accessibility classes; ε: contact potential; MJ: Miyazawa and Jernigan's potential.

given conformation c indeed have c as their unique ground state. More precisely, we generated 20 sequences by Gibbs sampling for 60 randomly chosen structures [see Additional file 8], i.e. 1,200 sequences for each potential, and performed a fold recognition experiment for the designed sequences, monitoring the score for the target fold using THREADER (Jones et al., 1992a) (figure 2.5 and table 2.I).

The solvent accessibility potential alone ($\alpha_{14ac} + \mu$, figure 2.5b) is not sufficient to provide specificity to the designed sequences, and behaves almost as poorly as the flat potential (μ, figure 2.5a). A mild improvement is seen when using the contact potential ($\varepsilon + \mu$, figure 2.5c): for 10% of the designed sequences the target fold is found among the best scoring folds (table 2.I), and the distribution of this ranking is skewed towards lower values. However, it is only with the combined potential ($\varepsilon + \alpha_{14ac} + \mu$, figure 2.5d) that a significant improvement is observed: for more than half of the designed sequences the target fold is found among the best 1% scoring folds, even though the average sequence identity with the native sequence is less than 10% in all cases (table

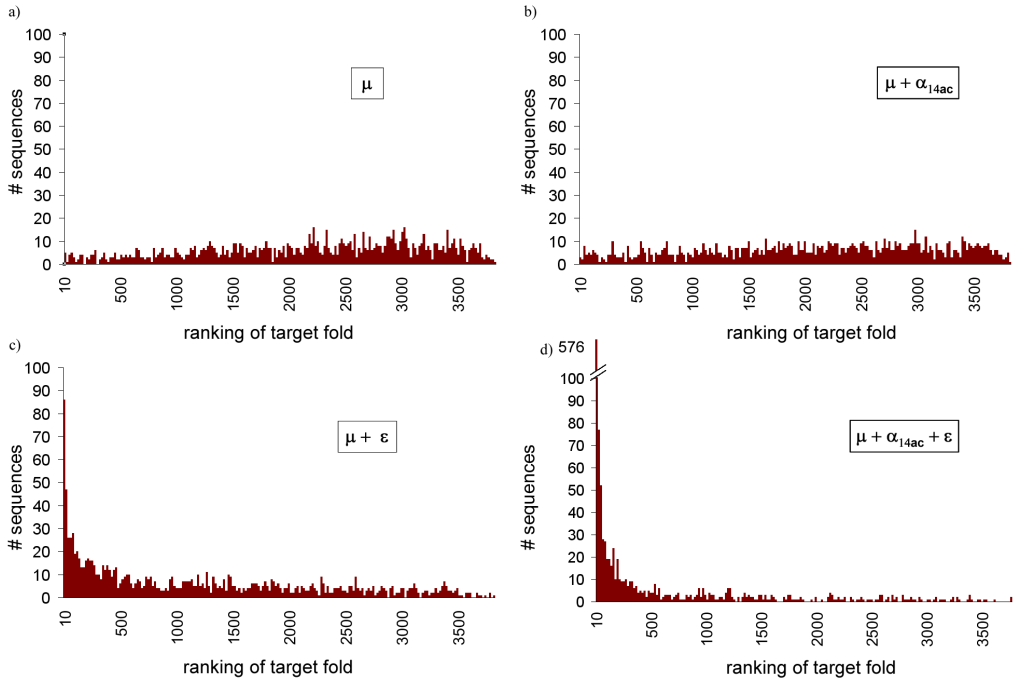


Figure 2.5: **Design specificity.** Histograms of the ranking of the target structure in a fold recognition experiment using THREADER. 20 sequences were generated for 60 randomly chosen structures, using (a) a flat (μ) potential, (b) a solvent accessibility, 14 classes ($\mu + \alpha_{14ac}$) potential, (c) a contact ($\mu + \varepsilon$) potential, and (d) the combined ($\mu + \alpha_{14ac} + \varepsilon$) potential.

2.I).

We also tested a subset of 120 randomly chosen designed sequences using another fold recognition program, LOOPP (Meller and Elber, 2001). LOOPP is based on a combination of several structure prediction methods, based on threading, secondary structure, sequence profile and exposed surface area prediction. The results obtained with this program were similar to those of THREADER: for 51.2% of the designed sequences using the combined ($\varepsilon + \alpha_{14ac} + \mu$) potential, the target fold was found as the first hit, and for 67.2% the target fold was found among the first 10 hits.

Table 2.I: **Specificity of designed sequences** Scores of a fold recognition experiment for designed sequences (see Methods). 1,200 sequences were sampled from $p(s \mid c, \theta)$ for each potential, and submitted to THREADER for fold recognition. Z-score ratio: Z-score of designed sequence / Z-score of native sequence in target fold.

Potential	Average Z-score ratio	SDev Z-score ratio	Ranking (median)	Target fold in top 1% (A)	Target fold in top 10%	Average seq. identity	Correlation between (A) and mean entropy/site
μ	-0.12	0.18	2249	0.5%	4.8%	5.76 %	-0.26
$\mu + \alpha_{14ac}$	-0.10	0.18	2090	0.4%	6.3%	6.65 %	-0.04
$\mu + \varepsilon$	0.13	0.16	816.8	10.7%	33.5%	6.69 %	0.23
$\mu + \alpha_{14ac} + \varepsilon$	0.45	0.23	32.7	53.6%	77.5%	7.82 %	0.64

In contrast, many of the current fold recognition programs based on sequence profile methods produced no significant hits (data not shown), which is not surprising, given that our sampling algorithm produces highly divergent sequences, with no similarity to any natural protein.

2.3 Discussion

The central idea of the present work is to reformulate the problem of devising statistical potentials for protein design as a statistical inference problem. This reformulation, based on the maximum likelihood (ML) principle, led us naturally to a gradient descent method, with the only additional aspect being that the gradient to follow is itself estimated by Monte-Carlo averaging.

The main advantage of this ML framework is that it guarantees an optimal predictive power of the resulting potential. In addition, it is very general, and can in principle be applied to any form of statistical potential. In particular, it is not restricted to coarse grained descriptions of proteins, and it could also be applied at the atomic level.

Interestingly, our gradient descent method turns out to be similar in spirit to an iterative scheme proposed by Thomas and Dill (1996a), although in that case the purpose was to optimize a potential in the context of the folding problem. Specifically, Thomas and Dill tune the potential so as to match the observed and expected number of contacts of each type, except that their expectation is taken on a set of alternative conformations, for a fixed sequence, whereas we take the expectation on a set of alternative sequences, on the conformation of interest. Note that Thomas and Dill derived their method from intuitive arguments, and not as a mathematical consequence of the ML principle.

These two alternative optimization schemes, obtained by normalizing either over the sequence or over the structure space, are quite distinct, at least conceptually. How the resulting potentials would differ in practice is more difficult to evaluate. Among other things, it will depend on how the approximation of lnZ_s based on the random energy model works. In the eventuality that it does not work well, it is likely that the contact term of our inverse potential will in fact combine two things: the information corresponding to the conformational energy of the sequence itself, which is also encoded in classical potentials optimized for threading, plus some information coming from the decoy term $\ln Z_s$. A way to settle this question would be to optimize a contact potential using, on the same learning set, both normalization schemes, and then compare the resulting values as well as their predictive powers.

2.3.1 Model assessment and comparison

The methodological framework proposed here offers reliable criteria for comparing the empirical fit of alternative models on real data. In this respect, it should be noted that the lack of a reliable objective criterion for evaluating different statistical potentials

has often been invoked for justifying the use of on-lattice idealized models (Mirny and Shakhnovich, 1996). However, on-lattice approaches are only moderately interesting, as they completely ignore the problem of the robustness of the learning method to model violation. Coarse-grained statistical potentials are by definition over-simplified models of proteins, and therefore, model violation is an intrinsic feature of the protein design problem. In this respect, the statistical language is interesting, since it is still valid, even for fitting and assessing models that are known to be imperfect.

On the other hand, the intuitive idea underlying cross-validation, i.e. measuring the rate of prediction of the native sequence, is quite simple, and has been invoked and used several times previously (Sun et al., 1995; Micheletti et al., 1998; Kono and Saven, 2001; Rossi et al., 2001; Jaramillo et al., 2002). What we propose here is a better formalization of this idea. Note that in contrast to previous methods, we do not measure the *marginal* native prediction rate at each site, but the *joint* probability of the native sequence. This can be important, as it accounts for possible correlations in the predictive distribution. For instance, two given positions may not display any particular pattern, when considered marginally, but may jointly follow charge or steric compensatory patterns. These phenomena will not be taken into account in the overall fit of the potential when measuring the marginal prediction rate, as is usually done. Technically speaking, the joint probability of the native sequence on the corresponding structure is extremely small, and cannot be evaluated just by counting the frequency at which the native sequence appears in the sample obtained by Gibbs sampling. For this, more elaborate numerical methods, such as thermodynamic integration, are required.

In the present case, the comparison between alternative models has allowed us to measure the relative contribution of each term of the potential and to refine the protein representation. The contact component turns out to be the most informative (figure 2.4), although it should be complemented with other energetic forms. Here, we have tested the addition of a solvent accessibility component, which significantly improves the fit of the model. Contact information and solvent exposure are correlated, which is reflected

in the fact that the fit improvement of each term is not additive.

Our model comparison method also gives us a direct way of choosing the optimal number of solvent accessibility classes (figure 2.3). Here, we found a number of 14 to 16 classes, which is higher than what one may have expected and than what is usually used. Note that this number depends on the way the classes are defined; here, the classes are based on quantiles, but as an alternative, we also tried a linear definition (evenly splitting the whole range of accessibility surfaces into D bins), which gave us an even higher optimal number of classes (20 classes, data not shown). In general, the present methodology could be used to investigate different definitions of accessibility classes, to refine the pairwise contact definition, or any other elements of the structure representation included in the potential.

The fact that our potential has a significantly better predictive power than that of Miyazawa and Jernigan (MJ, figure 2.4) is trivially expected, by construction of the ML potential. What is more surprising is that the MJ matrix is less fit than a simple solvent-accessibility profile. A possible explanation would be that Miyazawa and Jernigan's potential is based on the quasi-chemical approximation, which is now known to be somewhat drastic (Godzik et al., 1995; Thomas and Dill, 1996b; Skolnick et al., 1997), as it neglects correlations between observed pairing frequencies, due to chain connectivity and multiple contacts. Alternatively, it could mean that potentials optimized for folding are really not suited for protein design purposes. Testing other pairwise contact potentials, in particular those that do not rely on the quasi-chemical approximation (Tiana et al., 2004; Bastolla et al., 2001; Maiorov and Crippen, 1992; Tobi and Elber, 2000; Vendruscolo et al., 2000), would be a way to address this issue.

2.3.2 Sequence sampling

The method that we propose in this work is probabilistic in essence. As such, it offers a very natural framework for investigating the patterns induced by the models on distributions of sequences.

Specificity of the designed sequences

A sequence s designed for a target conformation c should not only be compatible with c, but also incompatible with competing folds. A rigorous solution to this problem involves a simultaneous search over the sequence and conformation space. It is possible, however, to achieve specificity without explicitly seeking to penalize competing states (*negative* design), if we rely on the approximation based on the random energy model, where the normalization constant of equation 2.4 can be considered as a function of the sequence composition only (Shakhnovich and Gutin, 1993; Koehl and Levitt, 1999). In our case, the normalization of the likelihood will also play an important role: since the total probability over all possible sequences has to be 1, maximizing the probability for a given sequence s_1 on its native conformation c_1 will lower the probability that another natural sequence s_2, with native conformation c_2, also gets a high probability on c_1. When many sequences are learnt in parallel, this phenomenon should ultimately favor specificity of s_2 on c_2, compared to all other conformations of the data set.

On the other hand, the extent to which the specificity is achieved will depend on the actual form of the potential used, as well as on the data base used for learning. To address this question, we produced a large number of sequences with four different potentials, and checked their ability to recognize the target fold, as measured by the Z-score ratio or by the ranking of the target structure in a fold recognition experiment. Indeed, an improvement of specificity is observed when using better potentials, suggesting that the method is effectively capturing specific dependencies between the conformation and the sequence of the proteins in the learning set, even for the simple forms of potentials tested here. For the combined $(\varepsilon + \alpha_{14ac} + \mu)$ potential, the average Z-score ratio of the designed sequences is similar to what has been reported for other protein design algorithms (Koehl and Levitt, 1999). Conversely, this also suggests that a more sophisticated potential may further improve the specificity of the sequences designed using our algorithm.

Conformation-dependent site-specific profiles

To compare natural protein sequences with those predicted by the optimized potentials, marginal, leave-one-out and empirical profiles (see methods) were generated for the 60 proteins used in the design specificity experiment described above; the profiles obtained for the best and the worst scoring structures are provided as supplementary materials [see Additional file 7]. Overall, leave-one-out profiles (figure 2.6a) and marginal profiles (figure 2.6b) do not display significant differences in the discriminative power between sites: the mean Shannon entropy per site is 0.743 ± 0.366 for marginal profiles, and 0.696 ± 0.428 for leave-one-out profiles. It is worth noting that the mean entropy per site for each protein , and the corresponding standard deviation, i.e. the average amount of information at each site and the variation between sites, are both correlated with the performance of the particular protein in the fold recognition experiment, and this, only for the combined ($\varepsilon + \alpha_{14ac} + \mu$) potential (table 2.I).

A detailed analysis of the leave-one-out profiles for a particular case, an alpha-aminotransferase, may be useful to understand which type of information is effectively captured by the potential, and which is not captured at all, thereby suggesting possible ways of improving the current form of potential.

First, regions of the protein that show little secondary structure (such as in positions 32-40, 55-65 and 82-88) contain less information (mean entropy per site = 0.756) than regions with local structure (mean entropy per site = 0.856). This is not surprising, since these regions typically have fewer contacts between residues, and thus the amount of information included in the protein representation is lower.

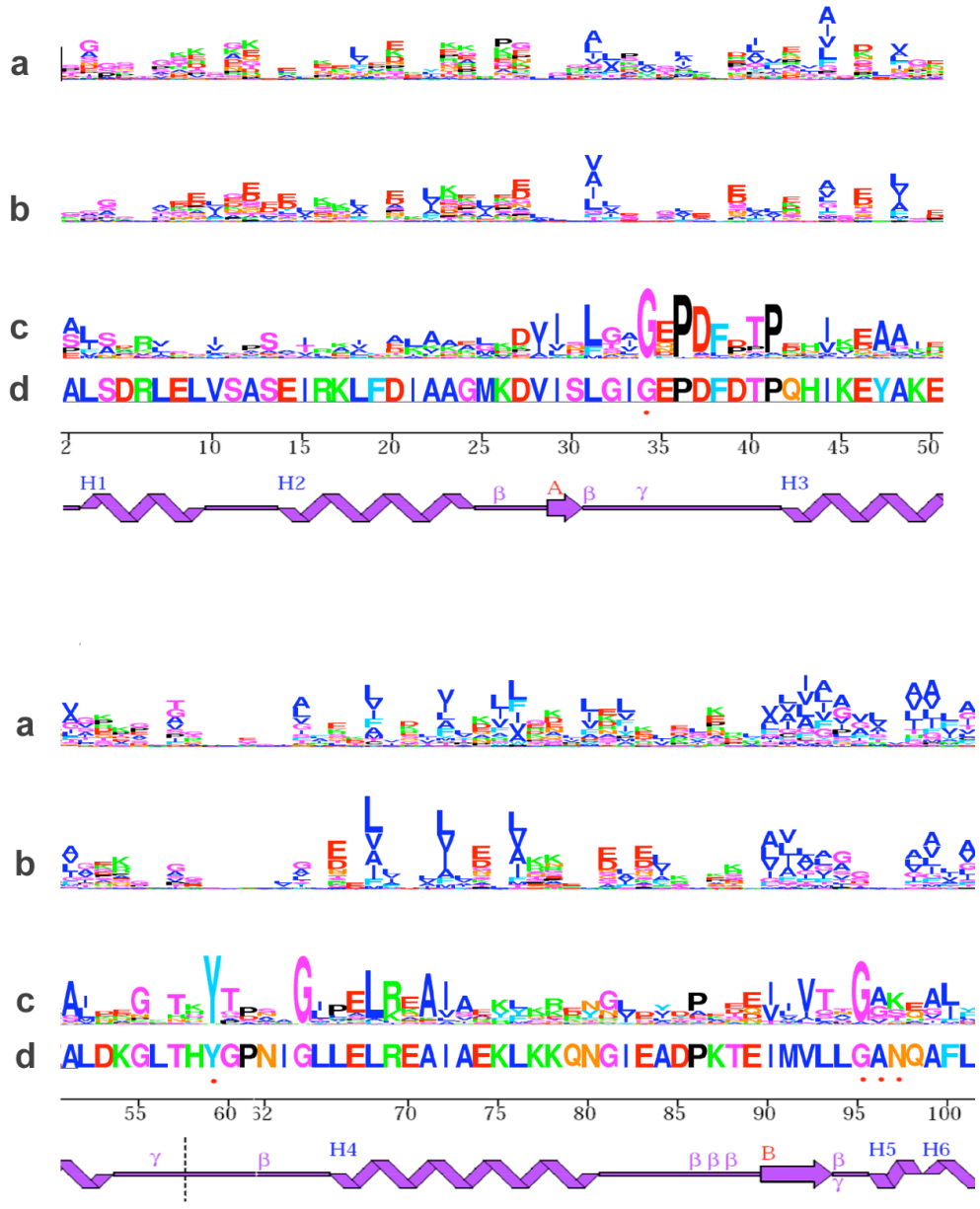


Figure 2.6: **Site-specific profiles.** Sequence logos of site-specific profiles induced on an alpha-aminotransferase ([PDB:1GDE], chain A), using a contact + solvent accessibility (14 classes) potential. From top to bottom: **(a)** marginal profiles, **(b)** leave-one-out profiles, **(c)** empirical profiles from a multiple sequence alignment of 162 sequences [see Additional file 4], and **(d)** native sequence of the reference protein. Secondary structure representation was taken from PDBsum (Laskowski et al., 2005). Red dot: residue interaction with ligand. Only the first 100 amino acids are shown; sequence logos for the whole protein are available as supplementary material [see Additional file 5][see Additional file 6].

Concerning regions with defined secondary structure, residue polarity is the information most easily captured. Charged residues are also distinctively inferred, as well as glycines, to a lesser extent (e.g. glycine 64, 81 and 95 – the latter predicted at position 94 or 95). In contrast, prolines are rarely correctly predicted, which is expected, since the properties most distinctive of prolines (such as phi-psi dihedral angles or local secondary structure) are not included in this particular form of potential.

Interestingly, some residues that have a crucial importance for the protein structure or function fail to be predicted, simply because the properties conferring their importance are not included in the protein description. This is the case of the amino acids that are in close interaction with a ligand (positions 34, 59, 96, 97).

Finally, the leave-one-out profiles display an interesting behavior with respect to positions where the amino-acid present in the reference sequence is not at all conserved in other members of the family. In some cases, they simply do not predict anything (e.g. glycines 24 and 60, or leucine 9, isoleucine 21, and alanine 23), which suggests that their limited importance in structure stability or function is recognized by the inverse potential. In other cases, the natural profile is even reproduced in the leave-one-out profile, instead of the amino acid of the reference sequence; such is the case for phenylalanine 100.

2.4 Conclusions

As illustrated by the sequence logos and the fold recognition experiments performed above, the predictive power of the models proposed here is encouraging, but nevertheless still weak. It is not yet clear to what extent this is due to the specific choice made concerning the form of the statistical potential, to the approximation of $\ln Z_s$ as a function of the sole composition of the sequence, or to yet other reasons. Most probably, we are facing a combination of several factors. The methods proposed here can now be used to address these difficult questions empirically.

In one direction, other approximations of $\ln Z_s$, less drastic than the random energy model, but still accessible in practice, can be investigated. For instance, following Deutsch and Kurozky (1996), the conditional probability of a sequence could be defined as:

$$p(s \mid c) \quad \propto \quad e^{-[E(S,C)-\langle E(S)\rangle]} p(s) \tag{2.37}$$

where the expectation $\langle \cdot \rangle$ is taken over a pre-defined set of decoy conformations. More sophisticated Monte Carlo methods, jointly sampling the sequence and conformation spaces, can also be imagined, in order to get more precise evaluations of $\ln Z_s$, while staying in the same global maximum likelihood formalism.

On the other hand, all the many statistical potentials that have been proposed over the last fifteen years may in principle be investigated in the same way as we have done here. In particular, distance-dependent potentials (Sippl, 1990) and main-chain dihedral angle potentials (Betancourt and Skolnick, 2004), which imply a richer representation of the protein structure, may result in models of greater predictive power. Other ways of implicitly considering side-chain conformation may also be easily incorporated into the model.

In a completely different perspective, it is possible to devise probabilistic models that are not exclusively defined in terms of a conformational free energy, even in a formal way. For instance, additional terms, concerning secondary structure aspects, interactions between successive positions along the sequence, or terms related to the folding constraints, can all be combined in an additive manner in the inverse potential. In fact, the model need not even be formulated in terms of a Boltzmann distribution, as long as the parameters are fitted by ML, and the predictive power of the resulting models is evaluated in a systematic way. Altogether, this amounts to setting up a robust statistical framework helping us to understand how, and to what extent, the sequences of natural proteins are determined by protein structure.

2.5 Methods

2.5.1 Structure representation

We used Miyazawa and Jernigan's definition of contacts (Miyazawa and Jernigan, 1985): each residue is represented by the center of its side chain atom positions; the positions of C^{α} atoms are used for glycine. Residues whose centers are closer than 6.5Å are defined to be in contact. The accessible surface of a residue is defined as the atomic accessible area when a probe of the radius of a molecule of water is rolled around the Van der Waal's surface of the protein (Lee and Richards, 1971). We used the program Naccess (Hubbard and Thornton, 1993) to make this calculation. When treating PDB files with multiple chains, solvent accessibility was calculated taking into account all molecules in the structure. The accessibility classes (percentage relative to the accessibility in Ala-X-Ala fully extended tripeptide) were defined so as to generate D equal-sized subsets of sites. The complete definition of accessibility classes is available as supporting material[see Additional file 1].

2.5.2 Monte Carlo implementation

In order to calculate the derivative of ω in the gradient descent procedure, expectations with respect to $p(S' \mid C, \theta)$ in equation 2.33 are evaluated numerically. A sample $(S_h)_{h=1..K_{EM}}$ drawn from $p(S \mid C, \theta)$ is obtained by a Gibbs sampling algorithm similar to that of Robinson et al. (Robinson et al., 2003). The elementary cycle of our Gibbs sampler is as follows: for each $p = 1..P$, and for each $i = 1..N_p$, each of the 20 amino acids is proposed at site i of protein p, by successively setting $s_i^p = a$, for all $a = 1..20$; in each case, the energy change ΔG_a induced by this point substitution is evaluated; then, s_i^p is set to amino acid a with probability $p_a \propto e^{-\Delta G_a}$. After Q cycles of burnin, a series of $h = 1..K_{EM}$ cycles are performed, and after each cycle, the current sequence, S_h, is

recorded. Once the sample is obtained, the expectation (2.32) is evaluated as

$$\langle \frac{\partial G}{\partial \theta} \rangle \simeq \frac{1}{K_{EM}} \sum_{h=1}^{K_{EM}} \frac{\partial G(S_h, C)}{\partial \theta} \tag{2.38}$$

and the derivative of ω with respect to θ follows immediately.

The overall gradient descent procedure runs as follows: we start from a random potential θ_0 and a random set of sequences, and perform the following iterative scheme:

- perform Q Gibbs cycles for the burnin, and K_{EM} additional cycles for the sampling itself. Keep the final sequences as the starting point of the next cycle.
- update θ by gradient descent, based on the estimate of the gradient obtained over the sample:

$$\theta_{n+1} = \theta_n - \delta\theta . \frac{\partial \omega(S)}{\partial \theta} \tag{2.39}$$

  where . is a scalar product, and $\delta\theta$ is a step-vector. In practice, the coefficients of $\delta\theta$ are tuned empirically, allowing three degrees of freedom, for the α, the ε, and the μ component of the potential respectively.
- iterate.

As a stopping rule, we monitor the evolution of $\omega(\theta)$ itself, which we evaluate every 100 steps by a numerical procedure (see below), and stop when $\omega(\theta)$ has stabilized. In practice, we used $Q = 100$ and $K_{EM} = 100$. At first sight, it would seem that a larger number of points K_{EM} would be needed to get a precise expectation, but in the present case one can rely on the self-averaging of the derivatives across the 100,000 sites of the database.

2.5.3 Likelihood evaluation

The difficult part in estimating the likelihood (or equivalently $\omega(\theta)$), for a given value of θ, is to obtain an evaluation of $\ln Y$. We do this by thermodynamic integration,

or path sampling (Ogata, 1989; Gelman, 1998), using the quasi-static method which we developed previously (Lartillot and Philippe, 2006).

First, for $0 \leq \beta \leq 1$, we define

$$G_\beta(s,c) = \beta \left(\sum_{1 \leq i < j \leq N} \Delta_{ij} \varepsilon_{s_i s_j} + \sum_{1 \leq i \leq N} \alpha_{s_i}^{v_i} \right) + \sum_{1 \leq i \leq N} \mu_{s_i}. \quad (2.40)$$

The associated probability distribution is:

$$p_\beta(s \mid c, \theta) = \frac{e^{-G_\beta(s,c)}}{Y_\beta}, \quad (2.41)$$

$$Y_\beta = \sum_{s'} e^{-G_\beta(s',c)}. \quad (2.42)$$

What we are looking for is $\ln Y_1$. As for $\ln Y_0$, it factors out, and can be computed directly:

$$\ln Y_0 = N \ln \left(\sum_{a=1}^{20} e^{-\mu_a} \right). \quad (2.43)$$

We can thus equivalently evaluate the difference $\ln Y_1 - \ln Y_0$. To do this, we rely on the following identity:

$$\ln Y_1 - \ln Y_0 = \int_0^1 \frac{\partial \ln Y}{\partial \beta} d\beta \quad (2.44)$$

$$= \int_0^1 \langle \frac{\partial G}{\partial \beta} \rangle_\beta \, d\beta, \quad (2.45)$$

where $\langle \cdot \rangle_\beta$ is the expectation over $p_\beta(s' \mid c, \theta)$.

In practice, the method consists in first equilibrating the Gibbs sampler at $\beta = 0$, and then, performing a series of $K_{Th} + 1$ cycles, where at each step, the value of β is increased by a small amount $\delta\beta = 1/K_{Th}$. The successive values of $\frac{\partial G}{\partial \beta}$ obtained during this quasi-static sampling scheme are recorded, and their average is our estimate

of $\ln Y_1 - \ln Y_0$:

$$\ln Y_1 - \ln Y_0 \simeq \frac{1}{K_{Th}} \left[\frac{1}{2} \frac{\partial G(s_0, c)}{\partial \beta} + \sum_{h=1}^{K_{Th}-1} \frac{\partial G(s_h, c)}{\partial \beta} + \frac{1}{2} \frac{\partial G(s_{K_{Th}, c})}{\partial \beta} \right]. \quad (2.46)$$

Note that these developments are for one protein, but the generalization over the database is straightforward.

In the conditions of the present work, $K_{Th} = 1,000$ is sufficient to obtain an estimate of $\ln Y_1 - \ln Y_0$ with an error less than one natural unit of logarithm.

2.5.4 Model comparison

We measured the fit of each model using cross-validation (CV): the potentials optimized on a first data set, i.e. the learning set, (θ_L) are applied on the second data set (the test set), and the log-likelihood is directly taken as a measure of fit. More precisely, for each model M,

$$CV_M = -\ln p(S_T \mid C_T, \theta_L, M), \quad (2.47)$$

where S_T and C_T are the sequences and structures of the test set. The difference with the CV score obtained for the flat potential (μ) is reported: $\Delta CV = CV_\mu - CV_M$.

2.5.5 Sequence sampling: site-specific profiles

Once an optimal value of θ is obtained, sequences compatible with a given conformation can be sampled from $p(s \mid c, \hat{\theta})$ by Gibbs sampling, and then further investigated. For instance, the frequency of each of the 20 amino acids (a) at each position (i) can be computed ($q_i(a)$) , yielding a vector of site-specific *marginal* profiles, graphically displayed as sequence logos (Schneider and Stephens, 1990). Alternatively, *leave-one-out* profiles can be obtained by computing the probability of each of the 20 amino-acids at each site of the test sequence, given the potential and the native sequence at all other

positions:

$$p(s_i = a \mid s_j, j \neq i, \theta). \tag{2.48}$$

We measured the amount of information displayed by the profiles using the site-specific Shannon entropy:

$$h_i = -\sum_a q_i(a) \ln q_i(a) \tag{2.49}$$

We compared both marginal and leave-one-out profiles to the *empirical* profiles, i.e. profiles displayed by natural sequences. We generated these empirical profiles from multiple sequence alignments obtained from the ConSurf-HSSP database (Glaser et al., 2005).

2.5.6 Sequence sampling: Design specificity

As a test for specificity, designed sequences were submitted to a fold recognition experiment, using the fold recognition program THREADER (Jones et al., 1992a). In THREADER, the compatibility of a sequence s for a given structure c is measured by the Z-score:

$$Z = \frac{\langle E(s,C) \rangle - E(s,c)}{\sigma} \tag{2.50}$$

where $\langle E(S,C) \rangle$ is the average of the THREADER statistical potential over all conformations of the decoy set, and σ is the corresponding standard deviation.

We randomly chose 70 structures of sizes ranging from 100 to 300 residues from the default THREADER dataset [see Additional file 8]. Structures whose native sequences produced a Z-score < 3 were discarded for the analysis. For each structure, c, we sampled 20 sequences from $p(s \mid c, \hat{\theta})$ by Gibbs sampling. These designed sequences were then submitted to THREADER (Jones et al., 1992a), and their specificity for the tar-

get structure c was measured by the ranking of c among all other structures, sorted by increasing Z-score.

A subset of 120 among the 1,200 sequences generated with the combined ($\varepsilon + \alpha_{14ac} + \mu$) potential (3-5 sequences for 23 distinct conformations, chosen at random; [see Additional file 8]) were also submitted to another fold recognition program, LOOPP (Meller and Elber, 2001), and the presence of the native conformation c as the first hit or in the first 10 hits was recorded.

2.5.7 Learning databases

We used proteins culled from the entire PDB according to structure quality (resolution better than 2.0 Å) and with less than 25% of mutual sequence identity [1] (Wang and Dunbrack, 2003). Two subsets of approximately equal size were obtained by partitioning the proteins randomly: DS1, 449 proteins, 100,077 sites, and DS2, 465 proteins, 99,894 sites. The final list of proteins is available as supporting material[see Additional file 2][see Additional file 3].

2.6 Additional Files

Additional files can be found on line at the url:

http://www.biomedcentral.com/1471-2105/7/326.

Additional file 1 — Extensive definition of accessibility classes

Additional file 2 — Data set DS1 - List of PDB identifiers of proteins used

Additional file 3 — Data set DS2 - List of PDB identifiers of proteins used

Additional file 4 — Multiple sequence alignment (Clustal format) used to generate sequence logos of figure 5.

1. In order to define the joint probability of the database as the product of the probabilities of the individual proteins, we need subsets of *independent*, i.e. unrelated, proteins. Of course, in practice, we have to make the approximation that highly diverged proteins are unrelated.

Additional file 5 — Marginal and leave-one-out profiles of complete protein partially displayed in figure 5

Additional file 6 — Empirical profiles of complete protein partially displayed in figure 5

Additional file 7 — Marginal and leave-one-out profiles of 10 proteins used in the design specificity experiment

Additional file 8 — List of PDB identifiers of proteins used in the design specificity experiment, and scores obtained for each one of the proteins, using the combined ($\varepsilon + \alpha_{14ac} + \mu$) potential

CHAPTER 3

STATISTICAL POTENTIALS FOR IMPROVED STRUCTURALLY CONSTRAINED EVOLUTIONARY MODELS

Preliminary to the work presented in the following chapter, we studied the effect that inherent biases on the training database have on the performance of the potentials, as measured by cross-validation scores (unpublished results). Briefly, we assembled training datasets based on different criteria, such as proportion of structured residues (alpha-helix or beta-sheet), size of the proteins, or size and number of ligands present in the crystallographic structure. Then, we compared the performance of the potentials optimized on these specific datasets with the performance of the potentials obtained using the general datasets of chapter 2. Of all studied variables, we found that only the secondary structure composition of training proteins has a significant impact on the resulting potentials, although the effect is much less drastic than the change in performance obtained by including additional structural elements. In other words, improving the functional form of the energy is much more important than avoiding biases in the training databases, at least at this stage of the work. We thus concentrated in this direction.

In the following article, the probabilistic framework presented in chapter 2 is used to optimize potentials of more complex functional forms. These potentials are then integrated into structurally constrained evolutionary models, and evaluated using the tools described exhaustively in Rodrigue et al. (2009). Several elements of the protein structure are considered, and evaluated by detailed statistical comparisons using large single-sequence data sets (by cross-validation), or three multiple sequence data sets in a phylogenetic context (by Bayesian approaches).

Molecular Biology and Evolution 2010, **27**(7):1546-1560; doi:10.1093/molbev/msq047

Statistical potentials for improved structurally constrained evolutionary models

Claudia L. Kleinman[1], Nicolas Rodrigue[2]Nicolas Lartillot[1] and Hervé Philippe[1]

1. *Département de Biochimie, Centre Robert Cedergren, Université de Montréal, Montréal, Québec Canada*
2. *Department of Biology, Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario, Canada.*

Keywords: protein structure, Bayes factor, statistical potentials, maximum likelihood, molecular evolution.

ABSTRACT

Assessing the influence of three dimensional protein structure on sequence evolution is a difficult task, mainly because of the assumption of independence between sites required by probabilistic phylogenetic methods. Recently, models that include an explicit treatment of protein structure and site interdependencies have been developed: a statistical potential (an energy-like scoring system for sequence-structure compatibility) is used to evaluate the probability of fixation of a given mutation, assuming a coarse grain protein structure that is constant through evolution. Yet, due to the novelty of these models and the small degree of overlap between the fields of structural and evolutionary biology, only simple representations of protein structure have been used so far. In this work, we present new forms of statistical potentials, using a probabilistic framework recently developed for evolutionary studies. Terms related to pairwise distance interactions, torsion angles, solvent accessibility and flexibility of the residues are included in the potentials, so as to study the effects of the main factors known to influence protein structure. The new potentials, with a more detailed representation of the protein structure, yield a better fit than the previously used scoring functions, with pairwise interactions contributing to more than half of this improvement. In a phylogenetic context, however, the structurally constrained models are still outperformed by some of the available site-independent models in terms of fit, possibly indicating that alternatives to coarse-grained statistical potentials should be explored in order to better model structural constraints.

3.1 Introduction

Protein structure has an undeniable role in shaping the evolution of protein coding sequences. Not only does the function of a protein depend primarily on the spatial arrangement of its atoms, but proper folding is crucial, since misfolded proteins tend to aggregate and cause unspecific cellular toxicity (Bucciantini et al., 2002; Dobson, 2003). As a result, over evolutionary time, protein structure changes much more slowly than the associated sequences (Flores et al., 1993; Russell et al., 1997). Despite this obvious role in evolution, the selective constraints imposed for maintaining a certain fold are still poorly characterized. The relationship between the structural importance of a residue and the purifying selection operating on that site is not straightforward, as several complex mechanisms may act simultaneously to accommodate variation. Natural proteins are more robust to random perturbations than expected by chance (Taverna and Goldstein, 2002a,b; Shakhnovich et al., 2005). They can accept substitutions at a large proportion of positions by small movements of interacting sites, or subtle shifts in the main chain conformation of spatially distant residues (Williams and Lovell, 2009), in addition to compensatory substitutions. Conversely, structural constraints are just one type of the many selective forces operating on sequences, which include maintaining specific function (such as binding and catalysis), folding kinetics, and regulatory constraints at the DNA and RNA level, to name a few. Disentangling the structural constraints from other constraints, from phylogenetic signal and from stochastic variation is a problem far from being solved.

One of the main difficulties for modeling evolution with explicit treatment of structural constraints is the site-interdependencies that the structure implies, which, for computational reasons, are handled by very few phylogenetic methods. Still assuming site independence, several attempts have been made to include an explicit treatment of protein structure (Overington et al., 1990; Wako and Blundell, 1994a,b; Koshi and Goldstein, 1995; Goldman et al., 1996; Thorne et al., 1996; Lio et al., 1998; Dimmic et al., 2000).

In all the cases, in addition to this important assumption, the evolutionary process is described as acting directly on amino acids, which has the shortcoming of confounding mutation and selection. More sophisticated models have been developed recently at the codon level (for a review, see Anisimova and Kosiol (2009) and Delport et al. (2009)), that permit the modeling of the interplay of mutation, selection and drift by making an explicit distinction between mutational and selective parameterizations. Among these, the structurally constrained models are of particular interest in our context. A statistical potential (a scoring system for sequence-structure compatibility) is used to evaluate the probability of fixation of a given mutation, assuming a coarse grain protein structure that is constant through evolution (Parisi and Echave, 2001). Robinson et al. (2003) combined this representation with statistical tools to make evolutionary inferences dealing with site interdependencies (Jensen and Pedersen, 2000; Pedersen and Jensen, 2001), establishing a model-based framework for assessing the effect of protein tertiary structure on evolution.

While adding the structural component to a given evolutionary model produces a substantial improvement in model fit (Rodrigue et al., 2006; Choi et al., 2007; Rodrigue et al., 2009), it is not sufficient to outperform state-of-the-art site independent models of codon substitution (Rodrigue et al., 2009). The oversimplified structural representation used so far in the sequence-structure compatibility scoring functions may play a central role in this issue. Due to the computational costs of the inference methods, a coarse grain representation of the protein is unavoidable; however, substantial improvement could likely be made regarding the form of the potentials in order to test more complex structural hypotheses.

Knowledge based potentials that yield reliable scoring functions while restricting the conformational search problem have improved over the last several years (Sippl, 1993; Miyazawa and Jernigan, 1996; Bastolla et al., 2000; Lazaridis and Karplus, 2000; Melo et al., 2002; Buchete et al., 2004; Boas and Harbury, 2007). They allow for variable levels of detail in describing the specific amino acid interactions and may account for

poorly understood physical phenomena, not exclusively related to protein stability (Boas and Harbury, 2007). However, the many potential functions developed in the context of protein structure prediction (where, given a sequence, a search is performed in the space of structures) may not be optimal for our purposes, since evolutionary studies pose the problem in terms of a protein design perspective: i.e. characterizing the set of sequences compatible with a given structure. The several approaches proposed in this direction are either based on lattice models (Chiu and Goldstein, 1998; Seno et al., 1998), or at the atomic level (reviewed in Boas and Harbury (2007). Besides implying heavier computational times, this latter representation has the problem of producing sequences too close to the particular native sequence, and implying a level of detail more difficult to reconcile with the assumption of a structure constant through evolution.

To overcome these limitations, we have recently developed a maximum likelihood framework for optimizing the parameters of a coarse-grain, residue level statistical potential, tailored for evolutionary studies (Kleinman et al., 2006; Bonnard et al., 2009). A pseudo-energy score $E(s,c)$ is defined as a sum of terms related to different structural descriptors, such as pairwise interactions or solvent accessibility. The probability of observing a database of sequences S given their native conformations C and the potential parameters θ, $P(S \mid C,\theta)$, is then maximized by gradient descent methods to obtain an optimal set of parameters. The method guarantees maximal predictive power for a given potential, and provides objective ways to selecting models for otherwise seemingly arbitrary definitions of the potentials.

In previous work (Kleinman et al., 2006; Bonnard et al., 2009; Rodrigue et al., 2009), a simple representation of the protein structure was used, consisting in a contact map supplemented with solvent accessibility information. In the present study, we aimed to model some of the the main protein structural features known to affect amino acid propensity: residue interactions, solvent accessibility, backbone conformation and flexibility of the residues. Residue interactions were described by replacing the binary contact map we previously used by distance-dependent pairwise interactions, the most

widely used representation for fold recognition and protein structure prediction (Jones et al., 1992a; Sippl, 1993; Jones, 1997; Xia et al., 2000). For describing backbone conformation, we focused on modeling torsion angles (Ramachandran et al., 1963; Kocher et al., 1994; Gilis and Rooman, 1997, 2001; Melo et al., 2002; Betancourt and Skolnick, 2004), or, alternatively, secondary structure conformation. Protein internal flexibility, in turn, critical for biological functions such as catalysis, allostery and interaction with other molecules, is a much more difficult feature to capture. Some information on protein dynamics is contained in the atomic displacement parameters (B-factors) of crystal structures, which reflect the fluctuation of atoms around their average position (Artymiuk et al., 1979; Frauenfelder et al., 1979; Sternberg et al., 1979). We included a term based on B-factors into the potentials, to assess the relevance of this measure as a surrogate for flexibility at the residue level. A cross-validation procedure, implicitly penalizing for model dimensionality, is used to evaluate the alternative combinations of these elements.

We will start by describing the derivation and validation of these new representations of the protein structure. Next, we incorporate them into a structurally constrained codon model of sequence evolution, and apply it to three protein datasets. We will discuss the selective constraints associated to these structural elements, and assess the performance of the new models against current site-independent models of sequence evolution.

3.2 Methods

3.2.1 Statistical potentials

3.2.2 Definition and optimization

Knowledge-based potentials are scoring functions that encode statistical patterns present in solved protein structures. They are inductive in nature, based on the idea that the propensity of an amino acid in a given site of a protein can be predicted by the observed frequency of that amino acid at other similar structural contexts in other proteins.

The probabilistic framework that we summarize below was used to optimize the parameters of different forms of statistical potentials by maximum likelihood, using non-redundant subsets of the PDB for training (Kleinman et al., 2006; Bonnard et al., 2009). Briefly, for a set of P unrelated proteins, each with a single associated structural conformation c^p and an amino acid sequence s^p of length N^p, let s_i^p be the amino acid at position i. Furthermore, assume that a model, M, consists of a set of structural contexts parameterized by θ, and that the observed frequencies of amino acids in each context can be modeled according to the propensity of each amino acid for that context using a Boltzmann distribution. The probability of obtaining a particular sequence is then (Kleinman et al., 2006):

$$p(s^p \mid c^p, \theta, M) = \frac{e^{-G(s^p|c^p,\theta)}}{Y^p}, \tag{3.1}$$

where $Y^p = \sum_{s'} e^{-G(s'|c^p,\theta)}$ is a normalization factor, taken over all possible sequences s' of length N^p, and $G(s^p \mid c^p, \theta)$ is the statistical potential. Adopting a Bayesian framework sampling parameters from their posterior distributions induces substantive computational complications, as the model leads to so-called doubly intractable distributions (Rodrigue et al., 2009). Instead, the parameters of the potential (for example, the contact energy for a given pair of amino acids) are estimated by directly maximizing the joint probability of the database:

$$p(S \mid C, \theta) = \prod_p p(s^p \mid c^p, \theta), \tag{3.2}$$

which can be seen as a likelihood. In practice, a leave-one-out pseudo-likelihood score function (Bonnard et al., 2009) was used in order to decrease the computational time of optimizations (for details, see Appendix 1, Supplementary Materials).

We will now focus on the definition of the statistical potential $G(s,c)$ (for simplicity,

we will omit the superscript p in the notation hereafter). It consists of two terms:

$$G(s \mid c, \theta) = E(s \mid c, \theta) - F(s \mid \theta). \tag{3.3}$$

The term $F(s \mid \theta)$ accounts for compositional effects, unrelated to the protein conformation. It cannot be solved analytically (Kleinman et al., 2006). Here, we use an approximation inspired from the *random energy model* (Shakhnovich and Gutin, 1993; Sun et al., 1995; Seno et al., 1998) and write:

$$F(s) = \sum_{a=1}^{20} n_a \mu_a, \tag{3.4}$$

where n_a is the number of occurrences of amino acid a in the sequence s. The unknown parameters μ_a represent the average propensities towards each amino acid, and are obtained in the optimization procedure along with all the other parameters.

$E(s \mid c, \theta)$, in turn, is the energy score. In our previous work (Kleinman et al., 2006; Bonnard et al., 2009; Rodrigue et al., 2009), $E(s \mid c, \theta)$ consisted of two terms:

$$E(s, c, \theta) = \sum_{i=1}^{N} \sum_{j=i}^{N} \Delta_{ij} \varepsilon_{s_i s_j} + \sum_{i=1}^{N} \alpha_{s_i}^{v_i}. \tag{3.5}$$

The first term is a contact energy: $\Delta_{ij} = 1$ if residues i and j are closer in space than a cut-off distance, and 0 otherwise, and ε_{ab} defines the contact energy between amino acids a and b. The second term encodes a solvent accessibility energy: for each residue, α_a^v represents the energy of amino acid a in the solvent accessibility class v, $a = 1..20$, and $v = 1..V$, where V is the total number of solvent accessibility classes considered.

In what follows, alternative definitions of $E(s \mid c, \theta)$ are explored, encoding different structural descriptors combined in a linear way:

$$\begin{aligned} E(s,c) =& \lambda_1 E_{Bfactor}(s,c) + \lambda_2 E_{torsion}(s,c) \\ &+ \lambda_3 E_{solv}(s,c) + \lambda_4 E_{dist}(s,c) \\ &+ \lambda_5 E_{ss}(s,c), \end{aligned} \tag{3.6}$$

where λ_i equals either 0 or 1, depending on whether the term is included or not in the potential under study. Although this linear formulation formally assumes independence between the terms, interactions between these elements do exist during the optimization, so that the parameters must be jointly optimized for each alternative functional form.

Several elements have to be determined a priori, such as the division of the parameter space into discrete classes, thus constituting a part of the model being assessed. The choice between alternative definitions was made based on model fit, measured by cross-validation (see below). Given the computational burden needed to incorporate site interdependencies into evolutionary models, there is a compromise to be considered in some cases, between the accuracy of the structural description and the computational cost of $E(s \mid c, \theta)$.

3.2.3 Model comparison and nomenclature

Alternative definitions of the structural elements considered yield different potentials, which can be interpreted as different models, and evaluated by standard statistical tools of model assessment. Here, once an optimal value of θ is obtained for each potential, the fit of alternative models is assessed by cross-validation (CV), consisting in training the potential on one dataset and calculating the log-likelihood score on a different, independent dataset. More precisely, for each model M,

$$CV_M = -\ln p(S_T \mid C_T, \theta_L, M), \tag{3.7}$$

where S_T and C_T are the sequences and structures of the test set, and θ_L are the parameters optimized on the learning set. The difference with the CV score obtained for a flat potential (μ, only accounting for compositional effects without any structural terms, i.e. $E(s|c) = 0$), normalized by the number of sites on the testing set N_T, is reported:

$$\Delta CV = \frac{CV_\mu - CV_M}{N_T} \tag{3.8}$$

We call the potentials obtained by the maximum likelihood framework *ML* potentials, and use the following abbreviations to refer to the structural terms included: *dist*, distance interactions; *cont*, contacts; *solv*, solvent accessibility; $Bfactor$, flexibility, measured by B-factors; *torsion*, main chain torsion angles; *ss*, secondary structure.

3.2.4 Main chain torsion angles

Backbone conformation can be described by the angle of rotation around the bonds of the main chain atoms, called the torsion angles omega, phi and psi. To capture the different conformation tendencies that different amino acids exhibit, we focused on modeling propensities for these angles. Torsion classes for angles phi and psi were defined based on a previously described version of the Ramachandran plot, which is divided into 9 discrete classes (Laskowski et al. (1996), supplementary figure S1 in Appendix II). For omega angles, on the other hand, two conformations were considered: *cis* or *trans*.

In this way, the conformation c of the protein includes the observed torsion class vectors T and W. The vector $T = (t_i)$ is the conformation of angles phi and psi associated with each site i, $t_i = 1...9$ and $i = 1..N$. The vector $W = (w_i)$, in turn, is the conformation of the angle omega at site i, with w_i being either *cis* or *trans*. The pseudo-energy associated with the three torsion angles has the following form:

$$E_{torsion}(s,c) = \sum_{i=1}^{N} \tau_{s_i}^{t_i} + \sum_{i=1}^{N} \eta_{s_i}^{w_i}, \tag{3.9}$$

where τ_a^t is the potential energy of amino acid a with angles phi and psi in conformation t, and η_a^w represents the potential energy for amino acid a with the omega angle in conformation w.

3.2.5 Secondary structure

As an alternative way of describing local structure, we derived a secondary structure potential:

$$E_{ss}(s,c) = \sum_{i=1}^{N} \varsigma_{s_i}^{l_i} \tag{3.10}$$

where ς_a^l is the energy parameter for amino acid a associated with the secondary structure element l. Secondary structure calculations were performed according to the method of Kabsch and Sanders (Kabsch and Sander, 1983; Laskowski et al., 1993). The ten elements considered are the following: residue in isolated beta-bridge; extended strand; 3/10 helix; alpha helix; pi-helix; bend; hydrogen-bonded turn; extension of beta-strand; extension of 3/10 helix; extension of alpha-helix. Alternatively, a simplified definition consisting of only 3 classes was also tested: alpha helix, beta strand and turn.

3.2.6 Flexibility of the residues

In order to capture some information about flexibility at the residue level, we implemented a potential based on the B-factor value at each site. B-factors were calculated either using alpha-carbons, or the average for all the atoms of the residue. Since the experimentally determined B-factor depends on elements such as the overall resolution of the structure, crystal contacts, and on the particular refinement procedures, B-factors from different structures need to be normalized before any comparison. We applied the

following normalization:

$$B_i^{norm} = \frac{B_i - \langle B \rangle}{\sigma_B} \tag{3.11}$$

where B_i is the B-factor recorded for residue i. σ_B and $\langle B \rangle$, in turn, are the standard deviation and the mean of B-factors for the given structure.

The energy score associated with B-factors has the form

$$E_{Bfactor}(s,c) = \sum_{i=1}^{N} \gamma_{s_i}^{g_i} \tag{3.12}$$

where γ_a^g represents the potential energy for amino acid a in the B-factor class g, $g = 1..G$. To determine the number of classes, G, several potentials were optimized with an increasing number of classes (from 0 to 50) and their fit was assessed by cross-validation. The classes were defined so as to generate G number of equal-sized subsets of amino acids (i.e. G quantiles) when analyzing 1000 randomly drawn proteins from the PDB.

3.2.7 Solvent accessibility

Solvent accessibility calculations were performed as described in Kleinman et al. (2006): the accessible surface of a residue is defined as the atomic accessible area when a probe of the radius of a molecule of water is rolled around the Van der Waal's surface of the protein. We used the program Naccess (Hubbard and Thornton, 1993) to perform this calculation, using the percentage relative to the accessibility in Ala-X-Ala fully extended tripeptide. When using PDB files with multiple chains, solvent accessibility was calculated taking into account all molecules in the structure. The optimal number of classes (in this case, 14) was determined by deriving potentials with an increasing number of classes, and evaluating their fit (Kleinman et al., 2006). We made the assumption that this optimal number of classes does not change when combining different structural terms, and verified that this was the case for the final form combining all the terms (data

not shown).

3.2.8 Distance-dependent interactions

The distance potential we implemented represents the separation of a pair of residues (in three dimensional space) as a discrete variable. An interval $R = [r_{min}, r_{max}]$ is defined, where r_{min} and r_{max} are, respectively, the minimum and maximum distance between two residues for considering an interaction. The interval is divided into D subintervals (also referred as classes) $r_d = [r^d_{min}, r^d_{max})$, $d = 1..D$, such that $r^1_{min} = r_{min}$, $r^D_{max} = r_{max}$, and $r^{d-1}_{max} = r^d_{min}$.

The distance x_{ij} between a pair of residues i and j is measured using either alpha-carbons, beta-carbons or the mass centers of the two side-chains. The energy term based on this distance has the form

$$E_{dist}(s,c) = \sum_{i=1}^{N} \sum_{j=i}^{N} \varepsilon_{s_i s_j}^{r_{ij}} \tag{3.13}$$

where r_{ij} is the distance class such that $x_{ij} \in r_{ij}$, and $\varepsilon_{ab}^{r_{ij}}$ defines the interaction energy between amino acids a and b in the distance class r_{ij}.

In order to define the intervals, i.e. specify D and the values of the different thresholds r^d_{min} and r^d_{max}, a preliminary analysis of the distribution of interactions between pairs of amino acids on 1000 randomly drawn PDB structures was performed. The region $R = [0,25]$ was partitioned into equal subintervals of 0.25Å. Let $f_r(a,b)$ be the frequency of observed interactions between amino acids a and b in the subinterval r, considered symmetrical, i.e. $f_r(a,b) = f_r(b,a)$. Let $f_R(a,b)$, on the other hand, be the frequency of interactions for the whole region 0-25Å. To compare these two distributions, the Kullback-Leibler divergence (KLD) was used:

$$KLD(f_r, f_R) = \sum_{a=1}^{20} \sum_{b=1}^{20} f_r(a,b) \log \frac{f_r(a,b)}{f_R(a,b)} \tag{3.14}$$

Note that KLD is always positive, and $KLD = 0$ when $f_r(a,b) = f_R(a,b)$.

3.2.9 Sequence sampling: site-specific profiles

Sequences compatible with a given conformation, induced by each one of the potentials, are obtained by Gibbs sampling as described in Kleinman et al. (2006) and displayed graphically as sequence logos. Profiles of natural sequences were generated from multiple sequence alignments obtained from the Consurf-HSSP database (Glaser et al., 2005). Alternatively, sequences were realigned using two programs, with default settings: Muscle (Edgar, 2004) and FSA (Bradley et al., 2009), producing essentially the same results. All the alignments are available as supplementary material.

3.2.10 Phylogenetic methods

3.2.11 Evolutionary model

Evolution of codon sequences is modeled as a Markov process defined in sequence space, fully determined by the matrix of instantaneous rates of change from one sequence (s) to another (s'). Mutation and selection are described as two separate processes, by the use of distinct sets of parameters. Following Robinson et al. (2003), selective constraints acting at the phenotype level are modeled by the statistical potential: the influence of the protein structure (a single conformation assumed constant along the entire tree) is represented by the difference in potential energy ΔG, with a parameter $\beta > 0$ modulating the strength of this influence. The parameters of $G(s \mid c)$ are fixed to the values obtained in the optimization by maximum likelihood described in previous sections. The model also includes an additional parameter ω, modulating nonsynonymous rates without regard to the amino acids involved.

The mutational specification, in turn, consists of two sets of parameters: $\rho = (\rho_{lm})_{1 \leq l,m \leq 4}$ is a set of symmetrical nucleotide exchangeability parameters, with $\sum_{1 \leq l < m \leq 4} \rho_{lm} = 1$; and $\varphi = (\phi_m)_{1 \leq m \leq 4}$ represents a set of global nucleotide equilibrium propensities, where

$\sum_{1\leq m\leq 4}\varphi_m = 1$.

In the complete model considered here, an off-diagonal entry of the Markov generator, corresponding to the instantaneous rate of substitution from s to s' is given by

$$R_{ss'} = \begin{cases} \rho_{s_{i_c}s'_{i_c}}\,\varphi_{s'_{i_c}}, & \text{if } \mathscr{A}, \\ \omega\rho_{s_{i_c}s'_{i_c}}\,\varphi_{s'_{i_c}}\,e^{-\beta(G(s')-G(s))}, & \text{if } \mathscr{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{3.15}$$

where

$\mathscr{A}$: s and s' differ only at the c^{th} codon position of the i^{th} site, and imply a synonymous change;

$\mathscr{B}$: s and s' differ only at the c^{th} codon position of the i^{th} site, and imply a nonsynonymous change;

and where s_{i_c} is the nucleotide at the c^{th} codon position of the i^{th} site of sequence s. Diagonal entries are given by the negative sum of off-diagonal entries in a given row. Note that when $\beta = 0$, the model is similar to the type of codon substitution model proposed by Muse and Gaut (1994).

As described in Rodrigue et al. (2009), the substitution process has a stationary probability given by

$$p(s^{\text{o}} \mid \theta, M) = \frac{1}{Z}e^{-2\beta G(s^{\text{o}})}\prod_{i=1}^{N}\left(\prod_{c=1}^{3}\varphi_{s^{\text{o}}_{i_c}}\right), \tag{3.16}$$

where Z is the normalizing factor:

$$Z = \sum_{s} e^{-2\beta G(s)}\prod_{i=1}^{N}\left(\prod_{c=1}^{3}\varphi_{s_{i_c}}\right), \tag{3.17}$$

with the sum being over all 61^N possible sequences.

We used the same priors and nomenclature as described in (Rodrigue et al., 2009).

We refer to the simplest model based on the mutational parameters only as MG, since it is inspired by Muse and Gaut (1994), and write MG-NS to refer to the model with a global nonsynonymous rate factor ω. When using the structurally constrained model based on the statistical potentials, we add the suffix ÐSC, giving MG-SC and MG-NS-SC. Finally, in the model referred as MG-NS$^{\mathrm{DP}}$, heterogeneity among sites is introduced by using a Dirichlet process as the law of the ω_i across sites (Huelsenbeck et al., 2006).

3.2.12 Bayes factors

Computational tools have been recently developed for sampling parameters from their posterior distribution under site-interdependent codon models and for the estimation of Bayes factors (Rodrigue et al., 2009):

$$B_M = \frac{p(D \mid c, M)}{p(D \mid c, M_{ref})} \tag{3.18}$$

where D represents the data, i.e. an alignment of nucleotide sequences related by a phylogenetic tree with a known topology, M is the sequence evolution model being evaluated, and M_{ref} represents the site independent model used as a reference (in the present case, MG).

Bayes factors are computed using thermodynamic integration, or *path sampling*, as described in (Rodrigue et al., 2009). In the case of the SC models, the procedure consists in sampling parameters using MCMC along a continuous path between M and M_{ref}, through a set of slight changes in the value of β. The result is a curve that represents a numerical evaluation of the fit of the model B_M as a function of β, the factor modulating the strength of the structural term in the evolutionary model (equation 3.15). The computations are made in duplicate, with different model-switch orientations, i.e. tracing the path from M to M_{ref}, and vice-versa, and we display both values obtained from these procedures.

Note that the evolutionary model proposed here imposes the same protein structure

(c) to all the sequences in the dataset, and that the particular native sequence corresponding to this structure (which we call s^c) is present in the alignment. In order to avoid the possible biases introduced by this presence, we can further decompose the marginal likelihood into two factors: one corresponding to the probability of the sequence state s^c, and another corresponding to the probability of observing all the other sequences (D^ϕ), conditional on s^c:

$$p(D \mid c, M) = p(D^\phi \mid s^c, c, M)\, p(s^c \mid c, M). \tag{3.19}$$

We then write

$$\begin{aligned} B_M &= \frac{p(D^\phi \mid s^c, c, M)}{p(D^\phi \mid s^c, c, M_{ref})} \frac{p(s^c \mid c, M)}{p(s^c \mid c, M_{ref})} \\ &= \left(B_M^\phi\right)\left(B_M^{s^c}\right). \end{aligned} \tag{3.20}$$

Formulated in this way, we are interested in distinct evaluations of two factors:

$$B_M^\phi = \frac{p(D^\phi \mid s^c, c, M)}{p(D^\phi \mid s^c, c, M_{ref})} \tag{3.21}$$

and

$$B_M^{s^c} = \frac{p(s^c \mid c, M)}{p(s^c \mid c, M_{ref})} \tag{3.22}$$

Given the reversibility of the overall substitution model, the factoring is arbitrary, but can be used to contrast contributions to model fit, with, for instance different leaf sequences taken for stationary probability factors.

The stationary probability factor, given in (3.16), can be computed for any leaf of the tree (Rodrigue et al., 2005), and, in particular, for s^c , making the calculation of the transient factor B_M^ϕ straightforward.

3.2.13 Datasets

3.2.14 Learning databases

We used proteins culled from the entire PDB according to sequence divergence in order to ensure independence (less than 25% mutual sequence identity), and to structure quality (resolution better than 2.0 Å) (Wang and Dunbrack, 2003). After discarding very small chains -less than 90 residues- subsets of 500 randomly drawn proteins were assembled. All datasets are available as supplementary material.

3.2.15 Phylogenetic datasets

Three datasets were used. The first, taken from Yang et al. (2000), consists of 17 vertebrate nucleotide sequences of the β-globin gene (144 codons). Structural information was extracted from the PDB file 4HHB. The second one, also from Yang et al. (2000), consists of sequences of the alcohol dehydrogenase taken from 23 species of *Drosophila* (254 codons) and the associated PDB file 1A4U. For both these datasets, we worked under the tree topology used by Yang et al. (2000). The third set consists of 34 calmodulin eukaryotic sequences, with a protein structure defined by the PDB file 1CFD and a tree topology estimated using phyML (Guindon and Gascuel, 2003) under the model JTT + F +Γ (Jones et al., 1992b; Yang, 1993). All datasets are available as supplementary materials.

3.3 Results and Discussion

3.3.1 Definition of statistical potentials and refinement of structural descriptors

The probabilistic framework described above was used to optimize the parameters of several forms of statistical potentials, based on different structural descriptors. These can be grouped in two types: pairwise interaction descriptors (contact map or distance-based matrix), and a series of site-independent components: solvent accessibility, tor-

sion angles, secondary structure and flexibility of the residues (table 3.I). As described in the Methods, the refinement of the structural descriptors is done by optimizing the alternative potentials and comparing their model fit in cross-validation experiments. We first analyze the site-specific terms, followed by the more complex site-interdependent descriptors.

Table 3.I: **Summary of class definitions used for the various elements of the optimized potentials**

Potential	Definition
$ML_{Bfactor}$: B-Factor	Average for all the atoms in a residue Normalized within each protein Five equal-sized classes
$ML_{torsion}$: Torsion angles	ϕ, ψ: 9 classes: - A a B b L l p X (Fig. S1) ω: *cis-trans*
ML_{solv}: Solvent accessibility	14 equal-sized classes (Kleinman et al., 2006)
ML_{dist}: Distance	Interaction center: side chain center Range considered: 3-11Å Resolution: 3-7Å interval: 0.5Å 7-10Å interval: 1Å 13 classes
ML_{cont}: Contact	Interaction center: side chain center Cutoff distance: 6.5Å
ML_{ss}: Secondary structure	10 classes (see Methods)

3.3.2 Site-independent descriptors

Aiming to capture flexibility at the residue level, we implemented a potential based on B-factor information. This measure was recorded either for the alpha-carbon or as the average for the whole residue, and normalized within each protein. A preliminary analysis on a large number of crystal structures shows that the distribution of B-factors is not identical for the different amino acids (figure 3.1(a) and S2(a)), indicating that this is likely an informative element. Moreover, the B-factors of particular regions in proteins seem to be conserved in protein families (Maguid et al., 2006), suggesting that

this measure correlates with a biological property. In order to define discrete categories for this feature, we analyzed the evolution of model fit as a function of the number of classes (figure 3.1(b)). When the number of classes increases, the fit of the model improves, until the penalization for model dimensionality starts to dominate the score. Not surprisingly, averaging the B-factor for all the atoms in the residue produced a markedly improved model fit compared to the alpha-carbon representation (more than twice the CV score, figures 3.1(b) and S2(a)).

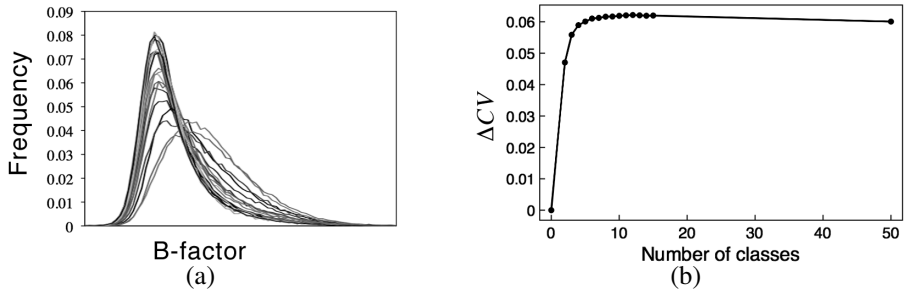


Figure 3.1: **B-factor terms** a) Distribution of B-factor for the different amino acids, in a nonredundant subset of PDB of 1,000 proteins. B-factor was calculated averaging B-factors of all the atoms in the residue, and normalized within each protein b) Evolution of cross-validation score of the potential as a function of the number of classes.

Backbone conformation, in turn, was described using either torsion angles or secondary structure. These two descriptions of the local conformation should in principle be redundant, with dihedral angles encoding richer information than the secondary structure, since they completely specify the position of the backbone. This is indeed reflected in our results. First, the torsion angle potential alone, $ML_{torsion}$, fits the data better (figure 3.2). Second, the contribution of the secondary structure term is less important for the combined potential $ML_{torsion,ss}$ (27% improvement with respect to $ML_{torsion}$, in contrast to the 55% expected if the terms were independent)(supplementary table S1, Appendix II). This reflects an important redundancy on the encoded information: for independent terms, one would expect approximately additive contributions to the fit of a combined model; conversely, completely correlated terms would produce a decrease in model fit

when combined, due to the penalization for model dimensionality. Considering different definitions of secondary structure (see Methods) produced only minor changes in the results (supplementary table S1, Appendix II).

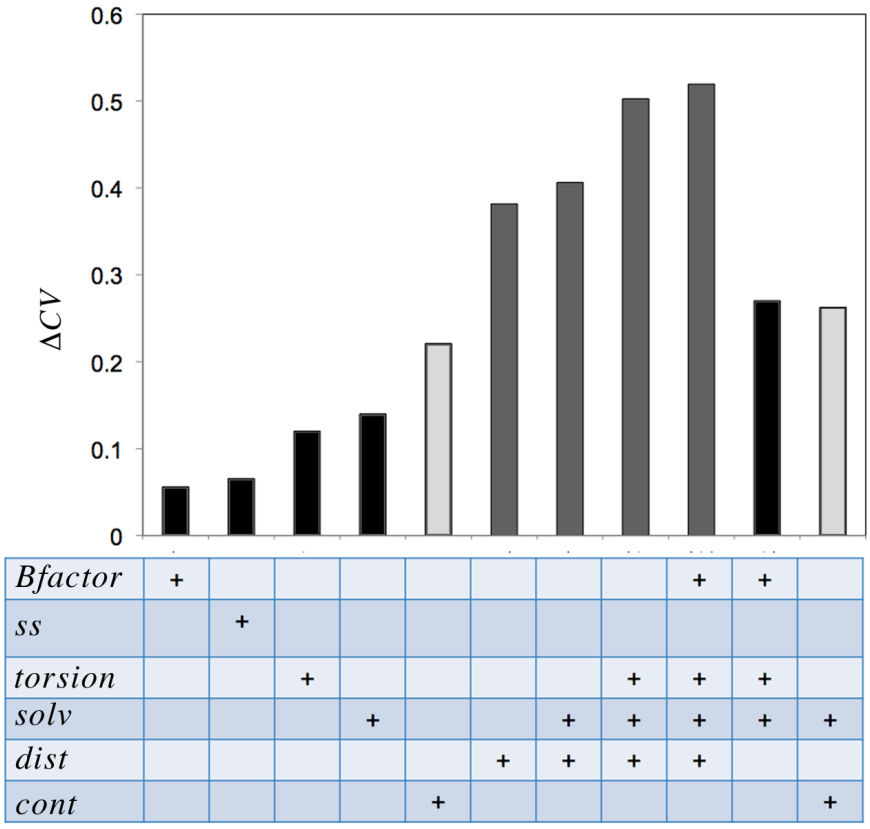


Figure 3.2: **Cross-validation scores for some of the different potentials obtained.** The average gain (relative to the CV score obtained with a flat potential, see Methods) for the 2-fold cross-validation experiment is reported. Black bars: site independent potentials. Dark grey bars: potentials containing distance-based terms. Light grey bars: potentials containing contact terms. The potentials were named according to the structural terms included in the definition: $Bfactor$, flexibility; *ss*, secondary structure; *torsion*, torsion angles; *solv*, solvent accessibility; *cont*, contact interactions; *dist*, distance interactions.

Of all the site-independent descriptors, the solvent potential, based on a discrete measure of the solvent accessible surface for each site, is the term producing the highest value of CV score. The optimal definition of this element was determined previously (Kleinman et al., 2006), in a similar way to the other terms described here, by optimizing

the alternative potentials and evaluating their fit. The good performance of this potential is not surprising, given the importance of hydrophobic interactions for stability and folding.

3.3.3 Pairwise interaction descriptors

The critical elements defining a distance-based potential are the choice of interacting centers, the range of distances considered, and the clustering of distance into discrete classes. In order to define these elements, we first performed an analysis of the distribution of pairwise interactions in known protein structures. Three interaction center definitions were successively considered: alpha-carbon, beta-carbon, and the center of mass of side chains. Ideally, in order to maximize the discriminatory power of the potential, distance classes should be defined in such a way that the distribution of interactions for each class is sufficiently different from the average distribution. In order to spot the areas where these distributions are distinctive, we partitioned the interval 0-25Å into small windows of 0.25Å and compared the 210 frequency vector of observed pairwise interactions in each window to the average distribution of interactions in the whole range, using the Kullback-Leibler divergence (*KLD*, figure 3.3(a)), for 1,000 randomly drawn PDB structures. Note that this is not meant as an optimization procedure, but as an heuristic method.

First, note the similarities in the overall shape of the plot for the three interaction centers studied. Windows corresponding to the shortest distances show the highest values of *KLD*, mainly due to sparse data and not because of a high amount of information in these regions. There is a peak at mid-range distances (around 6-7Å), and a small shoulder at longer distances (around 9-10Å). Not surprisingly, the value of *KLD* (which can be interpreted as the amount of relevant information) at these peaks correlates well with the level of detail of the corresponding structural representation. For the alpha-carbon representation, which encodes only information regarding the main chain, *KLD* is the lowest of the three. Using beta-carbon incorporates more information about the orien-

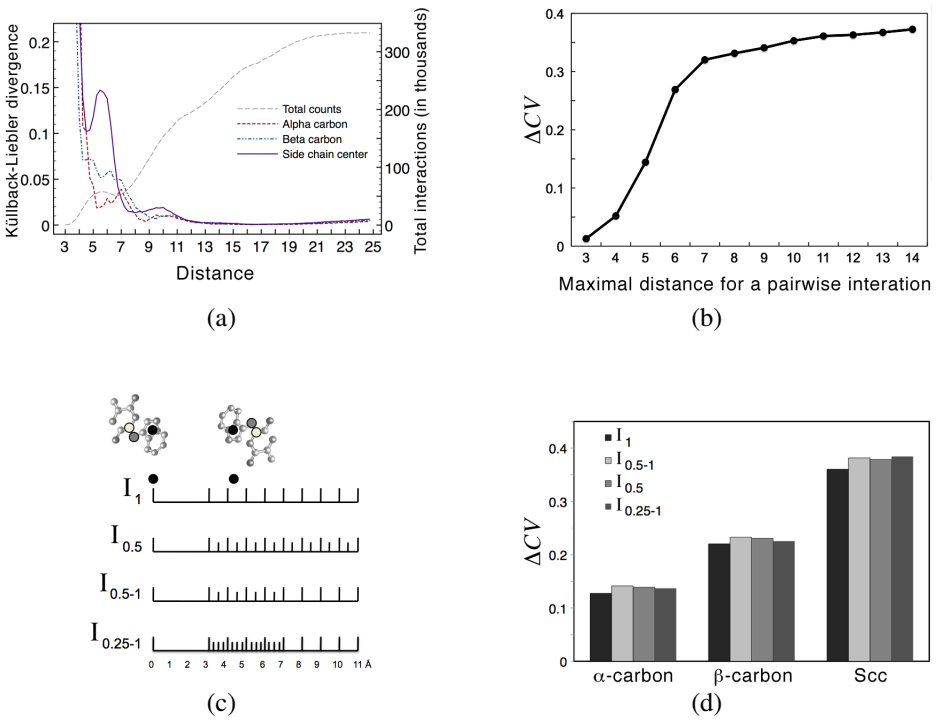


Figure 3.3: **Distance-based pairwise interactions.** a) The interval 0-25Å was divided in windows of 0.25Å, and the distribution of observed pairwise interactions in each window was compared to the average distribution in the whole region 0-25Å, using the Kullback-Leibler divergence (*KLD*). The total number of interactions, using side chain centers, for each window is shown (dashed grey line). b) Cross-validation score of distance-based potentials as a function of the distance range considered, using side chain centers. Distance intervals were partitioned in bins of 1Å. c) Graphical representation of the distance classes used in (d). The three interaction centers studied are marked with colored circles: black for side chain center, grey for beta-carbon, and white for alpha-carbon. Windows were defined as follows. The range 3-11Å was divided in windows of 1Å (named I_1; 9 classes), or 0.5Å (named $I_{0.5}$; 17 classes). Alternatively, the resolution was increased only for the interval 3-7Å, which was divided in windows of 0.5Å ($I_{0.5-1}$; 13 classes) or in windows of 0.25Å ($I_{0.25-1}$; 21 classes). D) Cross-validation scores of distance-based potentials as a function of the resolution and the interaction center used.

tation of the side chains, and consequently, *KLD* slightly increases. Finally, the highest peaks are found when using side chain centers for defining interactions.

Next, the *KLD* plot suggests an upper bound for the distances being considered: be-

yond 12Å, the distribution in each bin is indistinguishable from the general distribution, until around 21Å, where the distributions slowly start to diverge again. Not only is this divergence subtle, but including this region would imply an important increase in the cost of the calculation of $E(s,c)$, which is proportional to the number of contacts, approximately scaling with the volume of the sphere considered. It is known that for long distances, interactions are not residue specific, and are determined simply by solvation effects and the geometry of the molecule (Jones et al., 1992a), factors that will probably be modeled by other terms of the potential. Given the computational cost of incorporating site interdependencies into evolutionary models, we have a special interest in finding a range with few contacts considered, while remaining sufficiently accurate.

To further confirm the effect of the cutoff distance on the resulting potential, we derived several potentials by only varying their range, and dividing the resulting interval in bins of 1Å. The number of classes thus varies in each case, but the resolution and the interaction center are kept constant. The results obtained using side chain centers are shown in figure 3.3(b). The cross-validation score increases markedly when including distances corresponding to the high peak in the KLD plot (6-7Å). Adding the small peak at 9-10Å, however, has only a minor effect, indicative of some redundancies in these areas. A cutoff value of 11Å was used for subsequent analysis: increasing the range beyond such value does not produce a major improvement in the potential performance, but has the negative effect of drastically increasing the computational cost to calculate the energy.

Finally, we analyzed the effect of the resolution on the performance of the potentials. A scheme of the bins used is shown in figure 3.3(c). The region 0-11Å was considered. The interval 0-3 was not subdivided, given the small number of interactions it contains. The interval 3-11Å, in turn, was divided in bins of 1Å (named I_1), or 0.5Å ($I_{0.5}$). Alternatively, the resolution was increased only for the interval 3-7Å, divided in bins of 0.5Å ($I_{0.5-1}$), or 0.25Å ($I_{0.25-1}$). Increasing the resolution in the short distance interval ($I_{0.5-1}$) produces a better fit, for all the interaction centers considered (figure 3.3(d)).

For the potentials that use alpha-carbons or beta-carbons to describe an interaction, this is the optimal resolution obtained. This is not unexpected: potentials using a coarser description of proteins require a lower resolution for optimal performance, since over-parameterization penalties appear sooner. For all the interaction centers, increasing the resolution in the longer distance interval (7-11Å, $I_{0.5}$) was also detrimental (with respect to $I_{0.5-1}$, figure 3.3(d)), probably due to over-parametrization.

In principle, distance classes should be defined by maximizing differences not only with the general distribution of interactions, as we checked before, but also between different classes. We thus tested alternative discrete versions of the interval, not in a linear way, but based on the pairwise comparison of the *KLD* for all the different bins (supplementary figure S3, Appendix II). The performance of the potentials defined in this way was similar to the linear definition, suggesting that for this level of structural representation the resolution is already nearly optimal. No further work was thus done in this direction.

3.3.4 Combining the potentials

Figure 3.2 shows the cross validation scores for the potentials resulting from a linear combination of the terms described so far (table 3.I). As discussed before, the linear formulation of the combined potential $E(s,c)$ does not imply independence between the terms. Rather, it allows one to test for potential redundancies in the encoded information, by checking whether combined model configurations lead to interactions in terms of model fit.

It is worth noting that, when considering the potentials separately, the main improvement in model fit is brought about by the distance-based potential. It adds a considerable amount of information to the combination of all the site-independent descriptors, and performs better than the contact potential, solvent accessibility, or the combination of both that has been previously used (Kleinman et al., 2006; Rodrigue et al., 2009).

Solvent and pairwise interaction terms are highly correlated, and so the combined po-

tential $ML_{dist,solv}$ has a score merely 5% higher than the distance-based potential ML_{dist} (figure 3.2). On the other hand, this score is almost three times higher than the solvent potential alone ML_{solv}, suggesting that most of the information contained in the combined potential comes from the description of pairwise interactions.

Torsion angles, on the other hand, seem to encode orthogonal information to these two terms (figure 3.2 and supplementary table S1, Appendix II). This is consistent with the interpretation that they contain implicit information on the local conformation, independent of amino acid interactions, either with other residues or with the solvent.

As for the flexibility information encoded in the B-factor potential, although its inclusion produces a better fit than using a flat potential, this improvement is diluted when combining all the terms (figure 3.2 and supplementary table S1, Appendix II). The most plausible cause is a redundancy in the information encoded by the solvent accessibility and the flexibility terms; it is well known that residues in the core of proteins show less flexibility than those located on the surface, and the two measures are somewhat correlated (supplementary figure S2C, Appendix II). A similar behavior is observed for the secondary structure terms; the redundancies in this case are found with the torsion terms (as discussed above), and to a lesser degree with distance and B-factor terms (supplementary table S1, Appendix II).

The aim of this study being to incorporate the main factors affecting the protein structure, we restricted the analysis to a handful of terms whose importance is well established in the structural biology field. The model comparison and analysis of redundancies performed here, on the other hand, is general enough to be easily extended to other structural terms, or to terms not explicitly related to structural considerations.

3.3.5 Comparison of natural and designed sequences

Once the parameters of the potentials are optimized, we can perform an analysis in a protein design perspective by generating sequences from $p(s \mid c, \theta, M)$ by Gibbs sampling (Kleinman et al., 2006). The graphical display of these sampled sequences allows

for a qualitative analysis of the properties induced by the different potentials. An illustrative example is shown in figure 3.4, where the sampled sequences for a thioredoxin protein are contrasted to naturally occurring sequences.

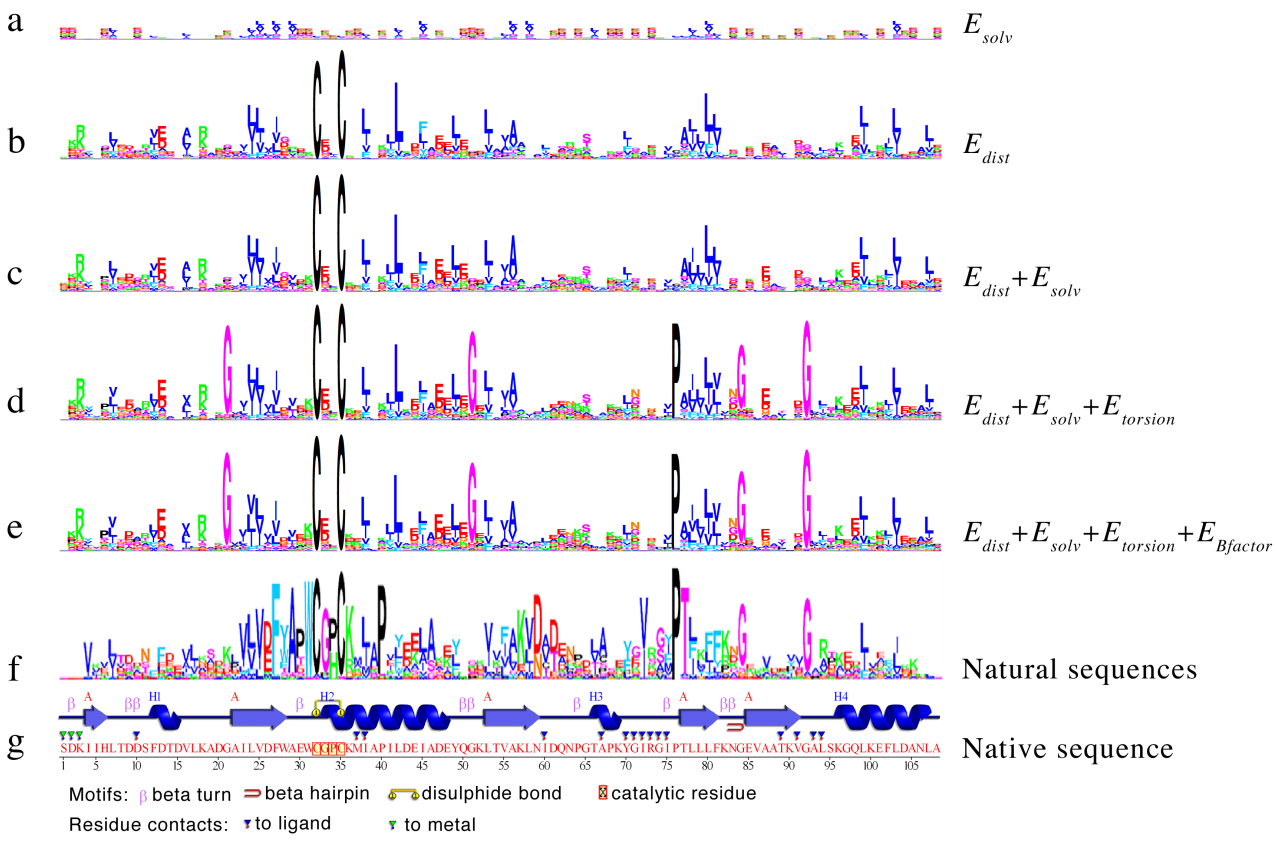


Figure 3.4: **Sequence logos of site-specific profiles** induced on a thiorredoxin (PDB: 2TRX, chain A), using the potentials a) ML_{solv}, b)ML_{dist}, c) $ML_{dist,solv}$, d) $ML_{dist,solv,torsion}$, e) $ML_{dist,solv,Bfactor,torsion}$. f) Profile obtained from a multiple sequence alignment of 162 eukaryotic sequences. g) Native sequence of the reference protein. Secondary structure representation from PDBsum (Laskowski, 2009). A color version of this image is available as supplementary material.

Note that the comparison performed here is not meant as a rigorous test of the performance of the potentials. Designed and naturally occurring sequences are conceptually different: while the former are free to explore the whole space of sequences compatible with the structure, the latter are constrained by their underlying phylogenetic structure.

Moreover, since the evolutionary relationship among the sequences is not accounted for when constructing sequence logos, the conservation observed in the natural profile is somewhat distorted by phylogenetic redundancy. Finally, natural sequences are highly diverged, and so the existence of many potential alignment errors cannot be dismissed.

Globally, designed sequences show a low degree of similarity to natural sequences. Residues that owe their conservation to known specific functional constraints are not predicted at all, as expected, simply because the properties conferring their importance are not being included in the protein structural description. Ligand binding sites (positions 10, 37, 38, 70-75, 89, 91, 93, 94), or residues in the catalytic site (positions 32-35), fall in this category. Apart from sites with known functional roles, the method fails to predict a number of conserved sites, particularly aromatic residues (positions 12, 27, 28, 31, 49, 81, 102) and specific polar interactions (e.g. Asp26-Lys82, Lys57-Asp61).

Nevertheless, a few general trends are apparent. Regarding the individual structural terms, distance-based potentials ML_{dist} predict very strongly disulfide bonds, and tend to predict mainly residue hydrophobicity (supplementary figure S4, Appendix II). The high redundancy between distance and solvent accessibility potentials suggested by the cross-validation experiments is also apparent here, as the sequence logos remain almost unchanged when adding the solvent terms. Several recent studies trying to link evolutionary rate to structural properties point to the solvent accessibility component as one of the main constraints (Goldman et al., 1998; Bustamante et al., 2000; Choi et al., 2006; Conant and Stadler, 2009; Franzosa and Xia, 2009; Gong et al., 2009). In all the cases, site independence is assumed. However, we can see that a rich description of pairwise interactions like the one presented here suffices to capture most of the information contained in the solvent accessibility terms, suggesting that the solvent exposure would not be in fact the main structural constraint.

A similar effect is observed for the B-factor information: it does not add any qualitatively different information, but it seems instead to modulate the strength of very few predictions (e.g. position 87). Torsion terms, on the other hand, provide new informa-

tion, changing the predictions for a few key amino acids such as prolines or glycines. In this particular example, thioredoxin has two prolines with very important structural roles. Pro76 is found in *cis* conformation, conserved through evolution and correctly predicted by the potentials including torsion terms. Pro40, on the other hand, produces a bending in a long alpha-helix; the latter feature is not currently modeled by the potentials, since the identity and conformation of neighboring sites is not considered when calculating the conformation of a residue, although it is known to affect the Ramachandran basin populations (Zaman et al., 2003). We are considering the inclusion of this feature in future work. As for glycines, potentials with torsion terms predict four of them very strongly; two of which (Gly84 and Gly92) are conserved in the profile of natural sequences, while the other two (Gly21 and Gly51) are not. However, this discrepancy is easily understood when looking at the actual alignment of natural sequences: both glycines are in fact present in more than one third of the sequences, but the alignment programs fail to position them properly because they are located in very divergent loops of the protein, where a high number of insertions and deletions are found.

Despite the limitations discussed above, a detailed analysis of the profiles of a particular protein like the one presented here allows for an intuitive visualization of the properties of the different statistical potentials. It spans a broad portion of the sequence space, using a large number of highly diverged sequences, which is more difficult to achieve within a phylogenetic framework.

3.3.6 Assessment in a phylogenetic context

Once the parameters of the potentials have been optimized, they can be inserted into a structurally constrained model of sequence evolution, and assessed in a Bayesian framework. The log-Bayes factors for two datasets of globular proteins, ADH and β-globin, are shown in table 3.II. The thermodynamic integration produces a curve representing the log-Bayes factor of each model as a function of β, the factor modulating the strength of the structural term in the evolutionary model (equation 3.15). This allows us, in ad-

dition to performing comparisons, to detect the optimal values of β for each model. We will first focus on this measure (table 3.II). Following the trend we observed using simpler SC models (Rodrigue et al., 2009), we find the optimal β to be positive, consistent with the case where sequences are selected for their compatibility to the structure. Note that the potentials were conceived to maximize a probability similar to the stationary distribution of the site interdependent codon model given in 3.16, although ignoring the contribution of the mutation bias, and with $\beta = 1/2$ (see Rodrigue et al. (2009) for details). The optimal value of β obtained is slightly below this expected value of $1/2$, maybe due to the fact that we are ignoring mutational pressure in the optimization procedure. Note that β-globin shows globally lower values of optimal β. This is probably due to the important structural features of this protein that are not described by the *ML* potentials considered here: the β-globin structure is greatly influenced by the prosthetic group, and by interactions with the other subunits of this oligomeric protein. In any case, for both proteins, models with richer structural description show a progressively higher optimal β: the better the structural representation, the stronger role this term plays in the evolutionary model.

The progression of the Bayes factor values when adding the structural terms one by one, similar to the trend observed before when measuring the fit of native sequence-structure pairs (figure 3.2), indicates that the sequence-structure patterns captured by the potentials are also meaningful in an evolutionary context. Once again, pairwise interactions are the most important single component contributing to model fit.

Although improving the description of the evolutionary process when contrasted to the MG model, the performance of the SC models remains altogether weak. MG-NS, a site independent model with only one global parameter modeling selection (ω), has a comparable performance (better in one case, worse in the other). Combining the structural specifications with the MG-NS model increases the model fit, though in a less important way than when adding them to a pure MG model. This is similar to what had been observed before (Rodrigue et al., 2009), which we interpret as a consequence of the

Table 3.II: **Natural logarithm of the Bayes Factor and optimal β for the models considered.** ω was included in the models either as a global parameter (noted as G), or with a Dirichlet distribution (noted as DP). Shaded cells show site-independent models of sequence evolution: MG-NS corresponds to row 5, and MG-NS^{DP} corresponds to row 10. MG was used as a reference model for the calculation of Bayes factors.

		ADH		β-globin	
ω	Potential	$\log B_M$	β	$\log B_M$	β
-	ML_{dist}	[145.90:146.01]	[0.383 : 0.390]	[81.99 : 82.16]	[0.312 :0.316]
-	$ML_{dist,solv}$	[162.35:162.82]	[0.390 : 0.392]	[90.00 : 90.03]	[0.325 : 0.327]
-	$ML_{dist,solv,torsion}$	[213.41:214.89]	[0.418:0.419]	[104.88 : 105.69]	[0.327 : 0.328]
-	$ML_{dist,solv,Bfactor,torsion}$	[222.37:222.76]	[0.414 : 0.419]	[114.47 : 114.64]	[0.331 : 0.333]
G	-	[316.3: 319.1]	-	[90.64 : 93.88]	-
G	ML_{dist}	[409.14 : 412.75]	[0.372 : 0.376]	[149.55 : 153.37]	[0.302 : 0.306]
G	$ML_{dist,solv}$	[417.96 : 421.10]	[0.370 : 0.381]	[155.69 : 159.04]	[0.297 : 0.317]
G	$ML_{dist,solv,torsion}$	[453.55 : 457.28]	[0.401 : 0.408]	[168.36 : 172.30]	[0.319 : 0.323]
G	$ML_{dist,solv,Bfactor,torsion}$	[458.32 : 461.92]	[0.397 : 0.399]	[174.73 : 178.90]	[0.325 : 0.326]
DP	-	[413.10 : 419.40]	-	[192.84 : 198.08]	-

overlap in the two approaches -ω and the SC settings- of modeling the purifying selection. Note, however, that despite this overlap, the combined MG-NS-SC model displays a fit that is in the order of MG-NS^{DP} (a site-independent model allowing heterogeneity of ω across sites), which the simpler SC models failed to attain before (Rodrigue et al., 2009). This suggests that the structural components of the model are explaining, if not the average nonsynonymous rate of substitution, a part of the heterogeneity of nonsynonymous rates across sites.

The mechanistic formulation of this approach allows for a simple interpretation of certain model violations. As an example, we analyzed a third protein, calmodulin, for which simple general rules of protein structure may not apply. Calmodulin acts as an intermediary protein that reacts to calcium levels and relays signals to numerous proteins. For this purpose, calmodulin undergoes major conformational changes (Hoeflich and Ikura, 2002). As such, this type of protein may not be well represented in the PDB. When applying the SC models, we observe a progressive increase in model fit (supplementary figure S5, Appendix II). However, this improvement is almost negligible

compared to the fit of MG-NS, which is five times higher. Consistently, neither layering the SC settings with the parameter ω, nor modeling heterogeneous ω parameters across sites with the MG-NS$^{\mathrm{DP}}$ model improve significantly the fit (less than 10% improvement). Since the global selective pressure in the present case is known to be unrelated to maintaining a single, rigid native structure, the detailed description of the amino acid interactions is not surprisingly meaningless in an evolutionary perspective.

3.3.7 Transient properties of the SC models

We also explored one additional aspect regarding the assessment of the SC models in this framework. Given our supervised learning procedure for optimizing the potentials, there is a risk of a bias towards the native sequence, i.e. the sequence that was used to obtain the crystallographic structure, a risk that increases with the level of detail in the structural description (Kuhlman and Baker, 2000). However, we are looking for a scoring function that predicts not only this native sequence s^{c}, but also more general sequence features that could be accepted by evolution under the particular structural constraints of c.

We can probably be confident that the coarse-grained modeling adopted here prevents such an overfitting, but this can be addressed quantitatively based on the following argument. We can decompose the Bayes factor into two factors (equation 3.20-3.22):

$$B_M = \left(B_M^{\phi}\right)\left(B_M^{s^{\mathrm{c}}}\right).$$

The factor B_M^{ϕ}, which we call the *transient* factor, measures the ability of the model to generalize beyond the native sequence, and predict new sequences related to the native one by their evolutionary history. The *stationary* factor $B_M^{s^{\mathrm{c}}}$, in turn, corresponds to the fit of the model on the native sequence itself. The results are reported in table 3.III. Note that both factors progress in the same order for the different potentials, and that the transient factor B_M^{ϕ} increases faster when enriching the SC model. This implies that

the structural specification is modeling meaningful selective constraints, and not merely describing too faithfully the relation between the native sequence and its structure.

Table 3.III: **Natural logarithm of the Bayes Factor and optimal β for the models considered,** considering separately the native sequence (s^c), and all the other sequences in the alignment (D^ϕ). See Methods for details. MG was used as a reference model for the calculation of Bayes factors.

	Potential	$B_M^{s^c}$	B_M^{ϕ}	β^{s^c}	β^{ϕ}
ADH	ML_{dist}	[76.84:76.86]	[69.85 : 69.92]	[0.413 : 0.417]	[0.356 : 0.360]
	$ML_{dist,solv}$	[85.27:85.40]	[77.33 : 77.86]	[0.410 : 0.414]	[0.372 : 0.373]
	$ML_{dist,solv,torsion}$	[106.92:107.28]	[106.49 : 107.61]	[0.416 : 0.418]	[0.420 : 0.420]
	$ML_{dist,solv,Bfactor,torsion}$	[110.83:110.99]	[111.40 : 111.92]	[0.418 : 0.419]	[0.412 : 0.419]
β-globin	ML_{dist}	[47.03 : 47.04]	[39.36 : 39.51]	[0.410 : 0.432]	[0.261 : 0.266]
	$ML_{dist,solv}$	[50.01 : 50.07]	[43.54 : 43.62]	[0.411 : 0.412]	[0.276 : 0.276]
	$ML_{dist,solv,torsion}$	[54.72 : 54.76]	[52.70 : 53.69]	[0.393 : 0.402]	[0.288 : 0.298]
	$ML_{dist,solv,Bfactor,torsion}$	[59.48 : 59.55]	[58.19 : 57.89]	[0.408 : 0.415]	[0.292 : 0.292]

Finally, note that the stationary factor represents an important contribution to the total Bayes factor, which may indicate that much of the model fit is obtained by explaining the native sequence. While it is true that, given that the model is time reversible, the marginal likelihood is invariant to the choice of s^c, the transient and stationary factors individually are not. In order to assess the role of the native sequence in this contribution, we repeated the experiment but considering all the sequences of the alignment, one at a time, as s^c (figure 3.5). We can see that the actual native sequence is not the one displaying the best stationary fit, indicating once again that the SC models are not merely predicting the native sequence. Changing s^c for other sequences produces relatively minor changes in the overall behavior of the plots for the two proteins tested (figures S6 and S7, Appendix II), suggesting that what is at stake here is a transient-stationary distinction rather than a native-non native one. The potentials have been optimized in a stationary state, without considerations related to the transient aspects of the evolutionary model; model violations may thus be more evident in the description of transient properties of the evolutionary process. A wide range of codon substitution models, presenting the

same associated stationary distribution to the one used here, but different transient forms, could be explored to further investigate this question (Thorne et al., 2007).

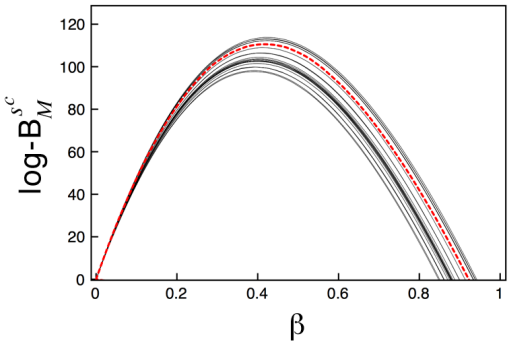


Figure 3.5: **Trace plots representing the stationary factor $B_M^{s^c}$ as a function of β**, the factor modulating the strength of the structural term in the evolutionary model. The computation was performed on the ADH dataset, using the potential combining torsion angles, solvent accessibility, pairwise interactions and B-factors ($ML_{dist,solv,Bfactor,torsion}$). In each curve, a different sequence from the alignment is taken as s^c. The dashed line corresponds to the case where the native sequence is taken as s^c.

3.4 Conclusion and perspectives

The main motivation behind this work is to incorporate explicit protein structure information in an evolutionary context, using a unified model-based statistical framework to assess the relevance of this information. To what extent are the factors known to affect protein structure -*in vitro*, in isolation and controlled laboratory conditions- shaping the evolution of protein sequences? Can we disentangle structural constraints from other selective forces? To address these questions, we derived statistical potentials with rich structural descriptions, optimized for evolutionary studies. We incorporated them into a structurally constrained model of sequence evolution, and evaluated them in a Bayesian framework.

We found that including detailed information on the protein structure improves the description of the evolutionary process. However, the performance of the potentials re-

mains relatively weak, compared to the most sophisticated site-independent models of evolution. Further improvements could be made regarding the specific form of the energy function, including terms related to interactions in torsion angles among successive positions along the chain (Betancourt and Skolnick, 2004), sidechain-backbone interactions (Buchete et al., 2004) or considering sequence separation ranges for distance interactions (Sippl, 1993). The modeling of flexibility, in particular, needs significant improvement. Even though B-factors have been previously used as an approximation of protein flexibility (Schlessinger and Rost, 2005; Yuan et al., 2005), our results do not support this role. The coarse grained representation of the structure provides an indirect way of allowing flexibility, but given its importance for protein function, an explicit modeling of this feature would be desirable. Other measures of protein dynamics could be explored, for example considering several conformations for each sequence in the learning database, each one representing different protein states, or homologous structures. In a different direction, refinements of the optimization procedure, which has not been modified here, should be considered, such as elements of negative design (Bolon et al., 2005), by the use of explicit decoy structures, or better approximations than the random energy model.

In any case, structural constraints represent only a fraction of the total selective constraint operating on sequences (Drummond et al., 2006; Pal et al., 2006; Drummond and Wilke, 2008). As shown by the logos of natural sequences, relatively few positions are strongly conserved, suggesting that the critical interactions for maintaining the overall structure may be relatively sparse. This has also been proven experimentally: a statistical function capturing coevolution in a sequence alignment, specifying very few key positions, suffices to produce correctly folded proteins *in vitro* (Suel et al., 2002; Socolich et al., 2005). Since Bayes factors are a global measure of how well all aspects of the data are explained by the model, if there are only a handful of positions constrained by the structure, the improvement in model fit will be minor.

More importantly, there is an intrinsic limitation of the modeling approach used here.

Statistical potentials are designed to capture general trends of amino acid propensities for average proteins, well represented in the learning dataset. However, as illustrated by the example of calmodulin, and to a lesser extent by β-globin, each protein structure has features critical for its function, folding and stability, which may be too particular to be accessible by estimating propensities over a large number of cases. Estimating the parameters for specific protein families, or, better yet, inferring them directly within the phylogenetic framework, along with the other parameters of the evolutionary model, may serve to overcome this limitation. In a more ambitious direction, more physically based representations and energy functions could be used to model protein structure, instead of relying on statistical potentials. This approach will certainly be computationally demanding, thus limiting the amount of data that can be analyzed, but it may prove to be a more direct and robust way to characterize structural constraints.

All in all, the quantitative analysis performed in this study, combining a mechanistic approach to modeling evolution with model-based statistical inference, may now be applied to study less well-characterized particular proteins, to answer more specific biological questions. In a different perspective, this framework can be extended naturally to handle other aspects of protein structure affecting sequence evolution, such as folding constraints, interactions with other proteins, or yet other phenotypic features, not exclusively related to the native conformation.

CHAPTER 4

ASSESSING THE INFLUENCE OF PROTEIN STRUCTURE ON SEQUENCE EVOLUTION: RELATIONSHIP WITH GENE EXPRESSION LEVEL

The following chapter presents a comparative study of the evolutionary process operating on proteins of very high and very low expression level. The structurally constrained evolutionary model developed in previous chapters is used to analyze two data sets, focusing on the estimated parameters related to selective constraints on protein structure.

The chapter is presented in the form of an article *in preparation*, which we have not submitted yet. We intend to extend the results presented here to a larger set of proteins before final publication. It is nonetheless included in this dissertation because I believe that the results obtained here are interesting on their own, and in the context of this dissertation, they provide a good illustration of how the framework we developed can be applied to a concrete biological question.

Assessing the influence of protein structure on sequence evolution: relationship with gene expression level

Claudia L. Kleinman[1], Nicolas Rodrigue[2], Nicolas Lartillot[1] and Hervé Philippe[1]

1. *Département de Biochimie, Centre Robert Cedergren, Université de Montréal, Montréal, Québec Canada*

2. *Department of Biology, Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario, Canada.*

ABSTRACT

We investigate selective constraints on the evolution of protein sequences imposed to maintain their three dimensional structure, and how such constraints are affected by protein abundance in *E. coli* cells. Structurally constrained evolutionary models, explicitly describing selection for protein structure, have recently been developed, as well as the statistical tools to objectively evaluate their performance. Because of their novelty and the high computational costs involved, analyses has been restricted thus far to a small number of proteins. Here, we apply this framework to a more representative set of proteins to study the variability of the structural influence on sequence evolution. We contrast patterns of model fit and posterior distributions of parameters associated with selection obtained on the most highly expressed genes with those of the lowest expression level. We find that there is a higher strength of selection for compatibility of the sequence with the folded state in highly expressed genes. More specifically, sequence changes that disrupt amino acid propensities for solvent accessibility classes are particularly penalized in abundant proteins, suggesting that this selective force plays a role in the long-standing correlation between evolutionary rate and expression level.

4.1 Introduction

Understanding the forces driving the evolution of protein sequences at the molecular level is a difficult problem, involving the modeling of several complex, overlapping and sometimes contradictory factors. The nature of these factors and their relative importance are among the open questions of evolutionary biology that, until recently, remained inaccessible to empirical studies. The traditional view has been that protein evolution is primarily determined by functional constraints and by the relative importance of the protein in the organism (Wilson et al., 1977). Systematic studies of evolutionary rates across genes, however, have shown that protein evolution is further affected by global factors such as genomic position of the genes, expression patterns or position in biological networks (Herbeck and Wall, 2005; Koonin and Wolf, 2006; Pal et al., 2006; Vitkup et al., 2006).

Perhaps surprisingly, one of the strongest negative correlates of a protein's evolutionary rate is its expression level: highly expressed genes evolve significantly slower than lowly expressed ones (Pal et al., 2001; Krylov et al., 2003; Drummond et al., 2005; Lemos et al., 2005). The association between expression level and sequence evolution is detected across a broad range of model organisms, though the correlation weakens for organisms of smaller population sizes, probably due to the lesser efficiency of natural selection in this case. An analysis of evolutionary rates in multidomain proteins (in which the domains are translated at the same rate) shows that they are substantially homogenized compared to the same domains in separate proteins (Wolf et al., 2008), supporting the hypothesis that protein abundance is one of the determinants of protein evolution, comparable with that of structural-functional constraints. Despite these correlational findings, however, the underlying causes of these observations remain a matter of debate.

It was recently proposed that the dominant cause of the covariation of expression level and sequence conservation rate is the selection for robustness to protein misfolding,

which is increasingly important for highly expressed genes due to the unspecific toxic effects of protein aggregation (Drummond et al., 2006; Wilke and Drummond, 2006; Drummond and Wilke, 2008, 2009). Computer simulations of protein evolution seem to indicate that the toxic effect of protein misfolding, indeed, could suffice to explain the observed correlation (Drummond and Wilke, 2008). Furthermore, highly expressed genes are less aggregation prone than genes of low expression level (Tartaglia et al., 2007, 2009).

Assessing the relative contribution, redundancy and synergistic effects of the different factors requires, however, more sophisticated modeling approaches. Most studies have focused so far on calculating the correlation between evolutionary rate and a particular protein feature, followed by statistical testing of the observed correlation. Multivariate techniques such as partial correlation or principal component analysis have been applied in several studies in order to dissect the relative importance of these factors, producing sometimes discrepant results (Drummond et al., 2006; Plotkin and Fraser, 2007; Kim and Yi, 2007).

In this context, an explicit mechanistic modeling approach linking aspects of phenotype to sequence change may help unravel the relative importance of these factors. Instead of using one model of evolution to calculate evolutionary rate and correlating this rate afterwards with a phenotypic trait, the protein feature under analysis is directly included in the modeling approach, thus reducing ambiguities when interpreting the results. Three dimensional protein structure is one such phenotypic trait. Recent progress in statistical tools have allowed its incorporation into standard probabilistic evolutionary models. A statistical potential (an energy-like scoring system for sequence-structure compatibility) is used to evaluate the probability of fixation of a given mutation, assuming a coarse-grained protein structure maintained constant through evolution (Parisi and Echave, 2001). This representation is combined with statistical tools to make inferences dealing with site interdependences (Robinson et al., 2003; Rodrigue et al., 2006), and to quantitatively evaluate their performance (Rodrigue et al., 2009). Given the computa-

tional costs involved and the relatively recent emergence of these tools, all the analyses have been thus far restricted to a very small number of proteins, with the exception of the work of Choi et al. (2007). In this case, numerous single-sequence datasets were used, with the Bayesian inference procedure based exclusively on the stationary distribution of sequences.

Here, we apply the structurally constrained modeling framework to two sets of proteins at both ends of the expression level range, to study how the selection for maintaining the three dimensional structure of a protein varies with proteins abundance. We find that there is an increasing strength of selection for compatibility of the sequence with the folded state and, more specifically, for meeting solvent accessibility requirements, suggesting that this particular selective force plays a role in the long-standing correlation between evolutionary rate and expression level.

4.2 Methods

4.2.1 Statistical potential

A statistical potential described in detail in Kleinman et al. (2010) is used to measure selective constraints related to the protein structure. Briefly, given a protein conformation, c, the potential, written as $G(s,c)$ for the pseudo-energy score of the amino acid sequence encoded by the codon sequence $s = (s_i)_{1 \leq i \leq N}$ is given by:

$$G(s,c) = E(s,c) - F(s,c) \tag{4.1}$$

The term $F(s)$ accounts for compositional effects, not necessarily related to the protein conformation, and is approximated by

$$F(s) = \sum_{a=1}^{20} n_a \mu_a \tag{4.2}$$

where n_a is the number of occurrences of amino acid a in the sequence s. The unknown parameters μ_a represent the average propensities of each amino acid. $E(s,c)$, in turn, is the energy score, implemented as a sum of four terms accounting for pairwise distance interactions, solvent accessibility, torsion angles and flexibility of the residues:

$$E(s,c) = E_{dist}(s,c) + E_{solv}(s,c) + E_{torsion}(s,c) + E_{Bfactor}(s,c) \tag{4.3}$$

The pairwise distance interaction terms have the form:

$$E_{dist}(s,c) = \sum_{i=1}^{N} \sum_{j=i}^{N} \varepsilon_{s_i s_j}^{r_{ij}} \tag{4.4}$$

where r_{ij} is the distance in space separating residues i and j, defined as a discrete variable, and $\varepsilon_{ab}^{r_{ij}}$ is the interaction energy between amino acids a and b separated by a distance r_{ij}. The other terms of the potential, which are site independent, take the form:

$$E_x(s,c) = \sum_{i=1}^{N} \alpha_{s_i}^{k_i} \tag{4.5}$$

where $x \in \{solv, torsion, Bfactor\}$, k_i is the corresponding structural class of site i, and α_a^k describes the propensity of amino acid a to be found in the structural class k.

Discrete classes for each term of the potentials were defined as described in Kleinman et al. (2010). Conformations *cis* and *trans* were considered for omega angles, while angles phi and psi were assigned to 9 regions of the Ramachandran plot. Flexibility at the residue level was modeled through average B-factor values normalized over each the protein, with 5 classes considered. Solvent accessibility measures were divided into 14 discrete categories. Finally, pairwise interaction terms were based on the distance between mass centers of side chains: the distance interval between two pair of residues was partitioned in discrete categories as follows: the range 3-11Å was considered; the

region 3-7Å was divided in intervals of 0.5Å, while the region 7-11Å was divided in intervals of 1Å.

The parameters of the potentials $\theta = \{\mu, \varepsilon, \alpha\}$ are derived from the statistical analysis of known protein structures, by maximizing the probability:

$$p(s|c,\theta) = \frac{e^{-G(s,c|\theta)}}{Y} \tag{4.6}$$

where $Y = \sum_{s'} e^{-G(s',c|\theta)}$ is a normalization factor, taken over all possible sequences s' of length N. This was done using gradient descent methods (Kleinman et al., 2006; Bonnard et al., 2009).

4.2.2 Evolutionary models

Evolution of codon sequences is modeled as a Markov process defined in sequence space, fully determined by the matrix of instantaneous rates of change from one sequence (s) to another (s'). Mutation and selection are described by two sets of parameters, that jointly define the substitution process.

Selective constraints acting at the phenotype level are modeled by the statistical potential $G(s)$ described before: the influence of the protein structure (a single conformation assumed constant along the entire tree) is represented by the difference in potential energy ΔG, with a parameter $\beta \geq 0$ modulating the strength of this influence. We will call $\Omega_{ss'}$ the term pertaining to the protein structure:

$$\Omega_{ss'} = e^{-\beta(G(s')-G(s))} \tag{4.7}$$

Note, as stated before, that the parameters of the potential function $G(s \mid c)$ are fixed to empirical values obtained in Kleinman et al. (2010) and are not considered as free parameters in the evolutionary model. On the other hand, following Robinson et al.

(2003) and Choi et al. (2007), a specific parameter β_x can be assigned to each structural term of the potential, and treated as a free parameter of the model:

$$\Omega_{ss'} = e^{-(\beta_{dist}\Delta E_{dist}(ss')+\beta_{solv}\Delta E_{solv}(ss')+\beta_{torsion}\Delta E_{torsion}(ss')+\beta_{Bfactor}\Delta E_{Bfactor}(ss')+F(s')-F(s)} \tag{4.8}$$

Finally, the model also includes an additional parameter ω, modulating nonsynonymous rates without regard to the amino acids involved.

The mutational specification, in turn, consists of two sets of parameters: $\rho = (\rho_{lm})_{1\leq l,m\leq 4}$ is a set of symmetrical nucleotide exchangeability parameters, with $\sum_{1\leq l<m\leq 4}\rho_{lm} = 1$; and $\pi = (\pi_m)_{1\leq m\leq 4}$ represents a set of global nucleotide equilibrium propensities, where $\sum_{1\leq m\leq 4}\pi_m = 1$.

In the complete model considered here, an off-diagonal entry of the Markov generator, corresponding to the instantaneous rate of substitution from s to s', is given by

$$R_{ss'} = \begin{cases} \rho_{s_{i_c}s'_{i_c}}\,\pi_{s'_{i_c}}, & \text{if } \mathscr{A}, \\ \omega\rho_{s_{i_c}s'_{i_c}}\,\pi_{s'_{i_c}}\,\Omega_{ss'}, & \text{if } \mathscr{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{4.9}$$

where

$\mathscr{A}$: s and s' differ only at the c^{th} codon position of the i^{th} site, and imply a synonymous change;

$\mathscr{B}$: s and s' differ only at the c^{th} codon position of the i^{th} site, and imply a nonsynonymous change;

and where s_{i_c} is the nucleotide at the c^{th} codon position of the i^{th} site of sequence s. Diagonal entries are given by the negative sum of off-diagonal entries in a given row.

Note that when $\beta = 0$, the model is similar to the type of codon substitution model proposed by Muse and Gaut (1994).

As described in Rodrigue et al. (2009), the substitution process has a stationary probability given by

$$p(s^{o} \mid \theta, M) = \frac{1}{Z} e^{-2\beta G(s^{o})} \prod_{i=1}^{N} \left(\prod_{c=1}^{3} \pi_{s^{o}_{ic}} \right), \tag{4.10}$$

where Z is the normalizing factor:

$$Z = \sum_{s} e^{-2\beta G(s)} \prod_{i=1}^{N} \left(\prod_{c=1}^{3} \pi_{s_{ic}} \right), \tag{4.11}$$

with the sum being over all 61^N possible sequences.

Note that the potentials have been devised to maximize a probability (equation 4.6) that is similar to the stationary distribution of the evolutionary model given in equation 4.10, with $\beta = 1/2$, although without considering nucleotide equilibrium frequencies. On the other hand, we know that the PDB proteins used to estimate the parameters of the potentials are also the result of an evolutionary process, and that their amino acid composition is thus influenced by the structure of the genetic code and by mutational pressures at the nucleotide level. A correction term was thus included in the term accounting for amino acid composition in the potentials (equation 4.2), to ensure consistency (for details, see Bonnard, 2010):

$$\mu'_{a} = \mu_{a} + \sum_{\sigma=\sigma_1\sigma_2\sigma_3|a} ln(\pi_{\sigma_1}\pi_{\sigma_2}\pi_{\sigma_3}), \tag{4.12}$$

where the sum is taken over all codon sequences σ coding for amino acid a. Equilibrium frequencies π_{σ} in equation 4.12 were estimated using the observed frequencies at the third codon position of four-fold degenerated codons in the proteins used for estimating the parameters of the potentials. In this way, both the structure of the genetic code and the nucleotide composition of PDB proteins are taken into account.

4.2.3 Priors and nomenclature

We used the same priors and nomenclature as described in (Kleinman et al., 2010). We refer to the simplest model based on the mutational parameters only as MG, since it is inspired by Muse and Gaut (1994), and write MG-NS to refer to the model with a global nonsynonymous rate factor ω. When using the structurally constrained model based on the statistical potentials, we add the suffix -SC, giving MG-SC and MG-NS-SC. Finally, in the model referred as MG-NS-DP, heterogeneity among sites is introduced by using a Dirichlet process as the law of the ω_i across sites (Huelsenbeck et al., 2006).

4.2.4 MCMC sampling

We used sampling techniques described elsewhere (see for example Lartillot and Philippe (2006); Rodrigue et al. (2008b)) that consist in drawing data augmentations conditional on parameters (and auxiliary variables) followed by updates on parameters conditional on the data augmentations. We used the approach described in Rodrigue et al. (2009) to draw data augmentations. Thermodynamic integrations for computing Bayes factors were performed as in Kleinman et al. (2010). General MCMC settings were tuned as described in Rodrigue et al. (2009): results are based on 10,500 cycles, removing the first 500 cycles as burn-in, and subsampling every 10th cycle, leaving 10,000 draws.

4.2.5 Datasets

We constructed two data sets according to protein abundance in *Escherichia coli*, using the measurements of Lu et al. (2007). Ten genes were selected from the most abundant proteins, and nine from the least abundant ones. The 19 genes fulfilled the following additional requirements: soluble, globular proteins, with at least one solved and published three dimensional structure, without large or numerous ligands. A minimum length of 150 residues was also required to ensure a sufficient phylogenetic signal. The

resulting list, as well as the associated PDB identifiers, are shown in Table 4.I.

For each gene, amino acid sequences of 50 gammaproteobacterial species (supplementary figure 4.8) were retrieved from GenBank, and aligned with Clustal (Larkin et al., 2007). The number of species in each alignment is sometimes less than 50, due to the absence of the gene in some of the bacterial genomes (table 4.I). DNA sequences were also retrieved from GenBank, and nucleotide alignments were reconstructed using the corresponding protein alignments as templates. Tree topologies were obtained with treefinder (Jobb et al., 2004), using the WAG+Γ model. Columns in the alignment without associated structural information in the corresponding PDB structure were removed, as well as positions having gaps in the *E. coli* sequence and positions with modified residues. This procedure eliminated in average less than 3% of the columns (table 4.I). Data sets were further reduced to limit Bayes factors calculation times to a week, eliminating the species with most missing genes (table 4.I and supplementary figure 4.8).

Structural information was extracted from PDB files using in-house developed libraries, as described in detail in Kleinman et al. (2010). For solvent accessibility, the accessible surface of a residue is defined as the atomic accessible area when a probe of the radius of a molecule of water is rolled around the Van der Waal's surface of the protein. We used the program Naccess (Hubbard and Thornton, 1993) to perform this calculation using the percentage relative to the accessibility in Ala-X-Ala fully extended tripeptide. When using PDB files with multiple chains, solvent accessibility was calculated taking into account all molecules in the structure.

4.3 Results

Two data sets of globular proteins were assembled, based on the level of expression of the corresponding genes and on the availability of a three dimensional structure (Table 4.I). The differences in length (average 323 and 391 sites for the high and low expressed proteins, respectively), as well as in number of species in the alignments, are not signifi-

cant (*p-value* 0.216 and 0.535, respectively). Nor did we find any significant differences in the proportion of structured residues for each protein, either as alpha helices (*p-value* 0.139), beta sheet (*p-value* 0.194), or the sum of the two measures (*p-value* 0.314). Most of the proteins in both data sets have enzymatic activities, without any major difference apparent at first inspection.

Table 4.I: **Datasets.**

Gene	Description	Uniprot	PDB	Length[1]	Nb. species[2]	Abundance[3]
tufB	Elongation factor Tu	P0A6N1	1DG1G	385(394)	33(33)	87672.53
gapA	Glyceraldehyde-3-phosphate dehydrogenase A	P0A9B2	1DC3A	308(330)	30(30)	49090.58
rpsE	30S ribosomal protein S5	P0A7W1	2AVYE	150(166)	46(46)	28222.73
serA	D-3-phosphoglycerate dehydrogenase	P0A9T0	1PSDA	404(409)	25(49)	23461.62
sodB	Superoxide dismutase [Fe]	P0AGD3	1ISAA	192(192)	43(43)	19637.41
icdA	Isocitrate dehydrogenase [NADP]	P08200	1SJSA	415(416)	27(27)	19479.77
rpoA	DNA-directed RNA polymerase subunit alpha	P0A7Z4	1BDFA	226(235)	42(42)	17803.54
fba	Fructose-bisphosphate aldolase class 2	P0AB71	1DOSA	358(358)	31(31)	16325.84
fabB	3-oxoacyl-[acyl-carrier-protein] synthase 1	P0A953	1G5XA	403(406)	25(49)	16042.49
pgk	Phosphoglycerate kinase	P0A799	1ZMRA	386(387)	25(50)	14681.75
sdhA	Succinate dehydrogenase flavoprotein subunit	P0AC41	2WDVA	588(588)	20(49)	114.05
aceB	Malate synthase A	P08997	3CUZA	529(532)	20(27)	116.68
pepP	Xaa-Pro aminopeptidase	P15034	1WLRA	440(440)	20(36)	131.52
metB	Cystathionine gamma-synthase	P00935	1CS1A	383(386)	25(30)	160.84
atpG	ATP synthase gamma chain	P0ABA6	1FS0G	219(230)	46(46)	182.82
nadB	L-aspartate oxidase	P10902	1CHUA	478(540)	20(47)	204.88
acpD	FMN-dependent NADH-azoreductase	P41407	1V4BA	197(200)	36(36)	216.43
asnB	Asparagine synthetase B [glutamine-hydrolyzing]	P22106	1CT9A	497(553)	25(34)	232.87
wcaG	GDP-L-fucose synthetase	P32055	1GFSA	317(321)	17(17)	238.73

[1] Length of the alignment, after eliminating columns without structural information. Number in parenthesis indicates original sequence length.

[2] Number of species used in the alignment. Number in parenthesis indicates the total number of species available that contain the gene.

[3] Estimates of absolute protein concentration per cell based on mass spectrometry assays, taken from Lu et al. (2007).

4.3.1 Bayes factors

In order to have a general picture of the behavior of SC models on the assembled data sets, we calculated, for each gene, the log-Bayes of four evolutionary models versus the simplest MG model (table 4.II). The values obtained for the MG-SC configuration are positive, indicating that a model including selection on protein structure is better describing the evolutionary process than a pure mutational model. Site independent models outperform the SC models in terms of model fit, however, evidence of the large fraction

of selective constraints unexplained by this coarse-grained structural model. Combining the structural specifications with the MG-NS model increases the fit of the model, which reaches values comparable to, but slightly lower than those obtained by a site independent model allowing for heterogeneity of selection across sites. This suggests that the SC model may be explaining, more than the global rate of evolution of the gene, the within-gene heterogeneity of the substitution process.

Table 4.II: **log-Bayes factors.**

Gene	MG-S-SC		MG-NS		MG-NS-SC		MG-NS-DP	
tufB	859	857	2020	2051	2389	2420	2595	2648
gapA	607	609	1467	1502	1795	1830	1902	1944
rpsE	276	278	1221	1259	1345	1384	1509	1548
serA	840	855	2482	2558	2892	2970	3177	3218
sodB	893	911	1794	1860	2185	2268	2351	2393
icdA	826	832	2625	2694	3049	2771	3233	3306
rpoA	565	565	1977	2060	2199	2282	2129	2357
fba	606	615	1859	2040	2171	2353	2302	2621
fabB	857	856	2465	2531	2854	2922	2969	3032
pgk	940	946	2362	2408	2814	2861	2731	2845
sdhA	1012	1015	2877	2963	3354	3442	3496	3550
aceB	992	1012	2561	2661	2950	3292	3652	3798
pepP	1088	1096	2118	2310	2613	2915	2906	3075
metB	843	846	2384	2540	2776	2940	3026	3193
atpG	469	469	1802	1838	2014	2049	2561	2639
nadB	985	996	2595	2744	3128	3281	3362	3500
acpD	820	872	1202	1345	1668	1837	1727	1832
asnB	962	965	3228	3425	3682	3880	3969	4143
wcaG	657	659	1526	1578	1741	2067	2066	2112

Previous analysis had been performed using a very reduced number of proteins (Rodrigue et al., 2009; Kleinman et al., 2010), or focusing on the stationary properties of the evolutionary model (Choi et al., 2007). Using a more representative set of genes and considering the substitution process as a whole, we were able to confirm the trends observed before. Bayes factors, however, do not allow us to directly address the question of how the influence of protein structure varies across different genes. Bayes factors, conceived to select models based on their ability to explain the same observed data, are

dependent on several elements specific to the data set under study, such as the length of the alignment, the number of sequences, and the total number of substitutions.

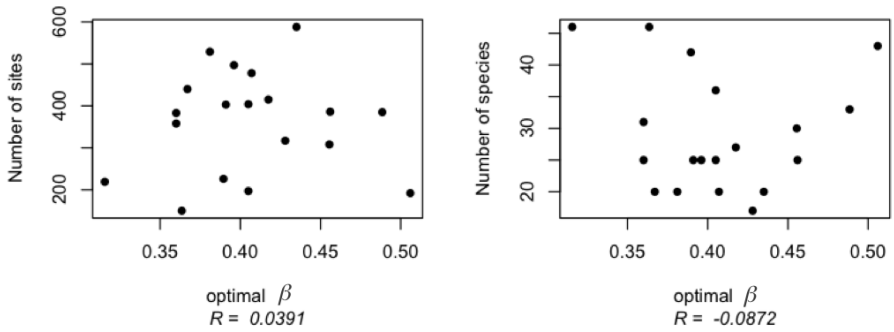


Figure 4.1: **Independence of optimal β and size of the dataset**, under MG-SC model. Left panel: optimal β plotted against alignment length. Right panel: optimal β plotted against the number of species in the dataset. Correlation coefficients are shown at the bottom of each plot.

4.3.2 Impact of the structural term in the evolutionary model

A more interesting parameter to analyze in order to compare the behavior of the SC models for different sets of proteins is the value of β, the factor modulating the strength of the structural term in the evolutionary model (equation 4.7). Thermodynamic integrations performed to obtain log-Bayes factors in the previous section allow for the determination of the β corresponding to the model of highest marginal likelihood, which we call optimal β. The potentials were designed to maximize a probability close to the stationary distribution of the site interdependent codon model given in equation 4.10, with $\beta = 1/2$. Thus, in the case where the model is accurately describing structural features, and where the proteins under analysis are well represented by the ensemble of PDB proteins, we would expect the optimal β to be found around this value. Lower values of β, on the other hand, indicate that the statistical preferences learnt over many different protein structures fail to account for the selective forces operating on the sequences being

tested.

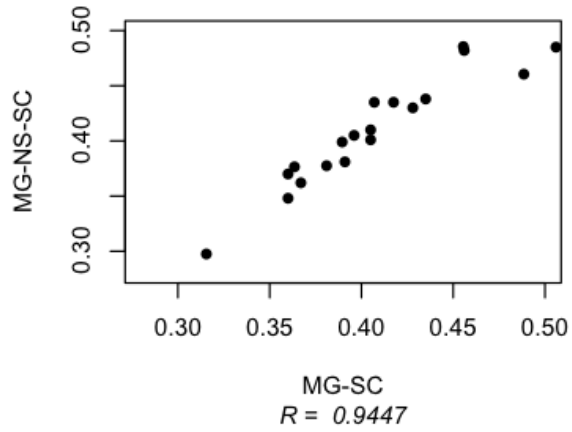


Figure 4.2: **Scatterplot plot of optimal β obtained under models with (MG-NS-SC) or without (MG-SC) the additional parameter ω.** Correlation coefficient is shown at the bottom of the plot.

As shown in figure 4.1, β seems to be independent of data set size, as measured by the length of the alignment or the number of species. Moreover, the values of β obtained with or without ω are well correlated (figure 4.2), suggesting that these two distinct modeling approaches of selection (SC specifications and ω) are not interfering at this level. In all other plots, then, only the results corresponding to the MG-SC model will be shown; the results obtained with MG-SC-NS are available as supplementary material.

In agreement with the hypothesis that highly expressed proteins are under different selective forces compared to the lowest expressed ones, and that this difference in selective constraints is related to the structural features of the molecule, we found the value of optimal β to be higher for the most abundant proteins (figure 4.3 and supplementary table 4.III). There is a positive correlation of β with expression level (figure 4.4), in particular for the most abundant proteins: the higher the expression level, the stronger the impact of the structural constraints on the evolutionary model. All in all, SC models seem to be better describing evolution for the most abundant proteins, suggesting that maintaining the global structure is more important for these molecules.

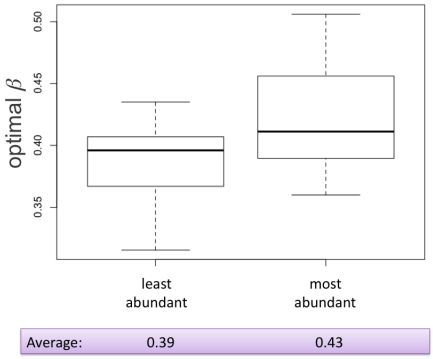


Figure 4.3: **Boxplots of optimal β for genes of high and low expression levels**, under MG-SC model. *p-value* = 0.104.

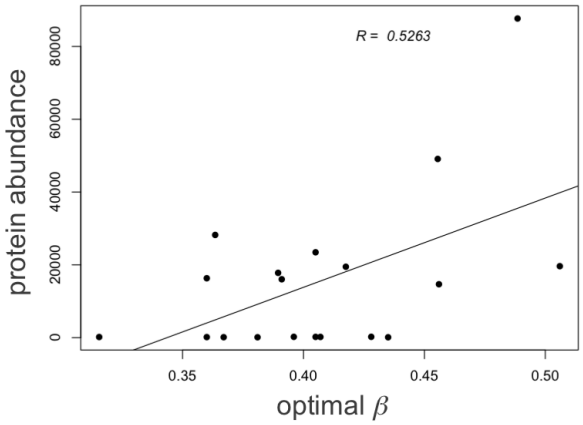


Figure 4.4: **Correlation of optimal β and protein abundance.** MG-SC model was used.

4.3.3 Solvent accessibility is under stronger selection than other structural elements in highly expressed proteins

In order to further investigate the differential selection operating on the most abundant proteins, we considered the following model (Robinson et al., 2003; Choi et al., 2007): instead of having a single parameter β modulating the stringency of the struc-

tural selection, a specific β_x parameter is associated to each structural element of the potential, with $x \in \{dist, solv, torsion, Bfactor\}$ (equation 4.8). In this way, the model can independently modulate the relative strength of each structural term. The β_x are considered as free parameters, endowed with uniform priors. We can then use MCMC to estimate the posterior distribution of each individual β_x.

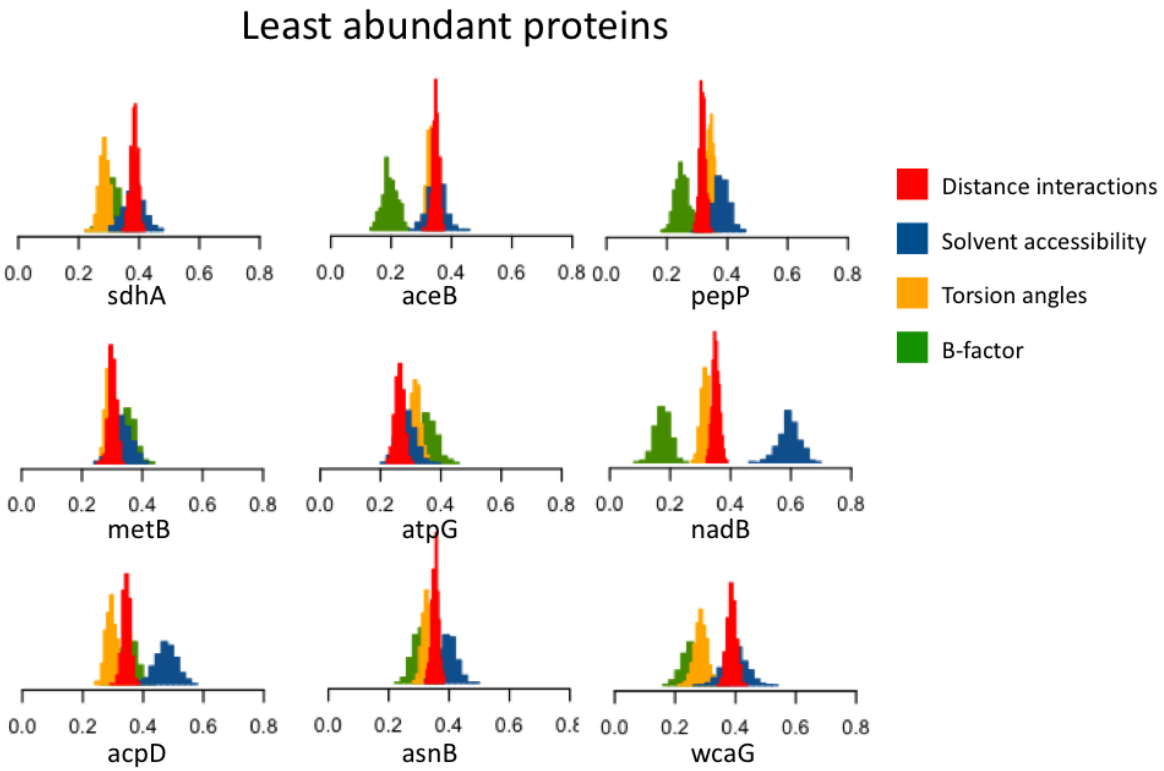


Figure 4.5: **Posterior distributions of β_x in least abundant proteins.**

The marginal posterior distributions over the β_x are shown in figure 4.5 for the least abundant proteins. In most cases, when analyzing individual genes, the distributions of the four β_x do not differ significantly, and their means are grouped around or slightly below the average value obtained for the global optimal β (0.39, see figure 4.3). An exception to this behavior is found in the gene *nadB* and, to a lesser extent, in *acpD*.

A different pattern is observed, on the other hand, for the most abundant proteins (figure 4.6). In this case, the general trend (7 out of 10 proteins) is that the distribution

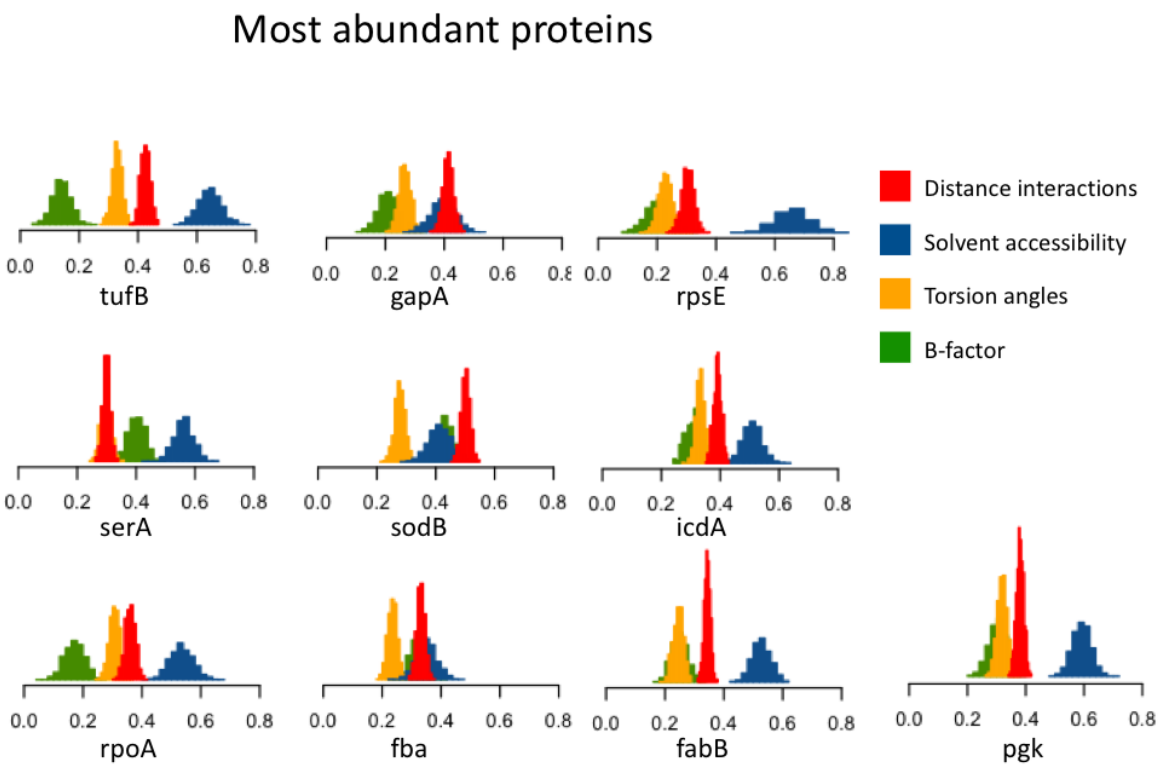


Figure 4.6: **Posterior distributions of β_x in most abundant proteins.**

of β_{solv} is significantly different from the other distributions, displaying a higher posterior mean. In many cases, the posterior mean of β_{solv} exceeds the expected value of 0.5, indicating that the weight of the solvent accessibility term when modeling evolution of a very abundant protein is higher than the weight this term has for an average protein found in PDB. Figure 4.7 summarizes the information contained in figures 4.5 and 4.6, contrasting the results obtained for the two sets of proteins side by side. We can see that β_{solv} in abundant proteins is significantly higher than all the other β_x, including β_{solv} in the lowest expressed proteins (*p-value* = 0.015). In other words, solvent accessibility requirements are specifically subject to stronger selection when a protein is highly expressed.

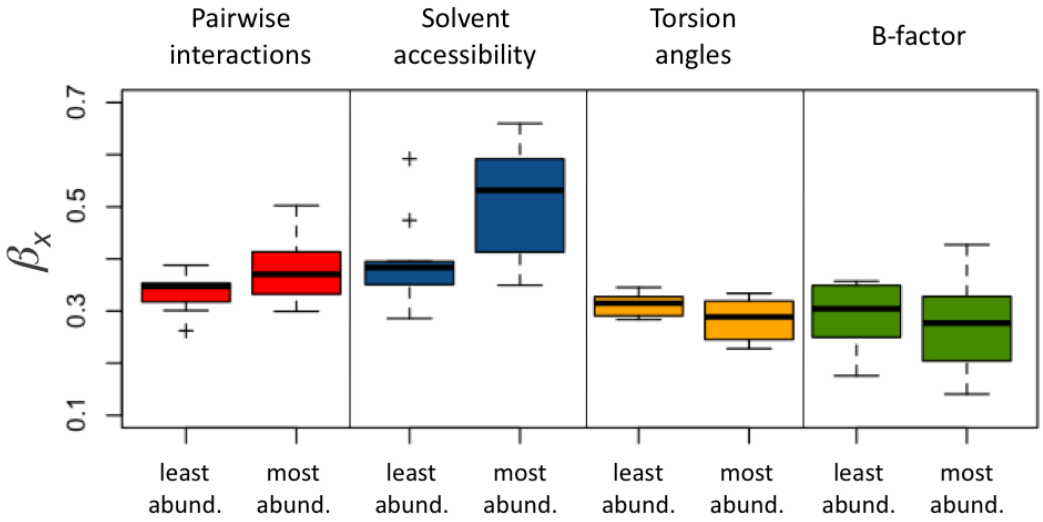


Figure 4.7: **Boxplots of posterior distributions of β_x.** The *p-value* for the comparison of β_{solv} in most and least abundant proteins is 0.015.

4.4 Discussion

We have made a comparative analysis of the evolutionary process acting on proteins of high and low expression levels, focusing on the role of three-dimensional protein structure. The approach presented here relies on modeling the microscopic effect of sequence changes on the organism fitness. By including the tertiary structure directly into the evolutionary model, the analysis is not limited to the evolutionary rate but implicitly considers other features of the substitution process, such as the nature of the amino acids allowed at each position and the site-dependences induced by the structure. Our results suggest that the more stringent selection currently found in highly expressed proteins is associated with a stronger selection for maintaining the folded state: for a given difference in potential energy (ΔG), the probability of fixation is lower if the mutation affects an abundant protein. More specifically, sequence changes that disrupt amino acid propensities for solvent accessibility classes are particularly penalized.

When a single global β parameter was used, the difference was, however, not significant (figure 4.3), which could be due to the small size of the data sets studied. Never-

theless, the result obtained when assigning a specific β_x to each structural term (figure 4.7) suggests that the reason may be that, with only one of the structural elements being subject to particular selective constraints, the signal is attenuated when considering a single global parameter modulating the structural term in the model. Extending the study performed here to a larger set of proteins should confirm our results and allow a better assessment of their significance.

Note that the measure of protein abundance used here corresponds to the amount of protein in a cell in standard laboratory conditions. On the other hand, for proteins whose expression level changes over the life time of the organism, selective constraints will operate if the effect of a sequence change is deleterious, even if the harmful effect lasts for only a short period of time. This could explain the posterior distributions of β_x observed for the lowly expressed gene *nadB*, which are similar to the ones observed for highly expressed genes: *nadB* has been shown to be up-regulated in some *Bacillus* species, during infection of host macrophages (Bergman et al., 2007), or in the presence of sulphate or methionine (Auger et al., 2002). In *E. coli*, it is predicted to be up-regulated in anaerobiosis (Schramm et al., 2007).

Although the results obtained here do not exclude additional forces differentially operating on highly expressed proteins, they specifically point to a structural component as a target of selection. In a similar direction, Cherry (2010) found that highly expressed proteins share compositional properties with thermophilic proteins, which tend to have more stable folds than proteins from mesophiles. Alternative constraints, covariating with expression level but not directly related to features of the three dimensional protein structure, should not affect the statistical fit or the estimated parameters of the SC models. Examples of these are amino acid biosynthetic cost, translational efficiency (Akashi, 2001) and other properties such as regulatory elements at the RNA level or position in biological networks. SC models rely on an explicit description of the effect of substitutions on the phenotype; the effects observed are thus conditional on the protein structure, and not a mere correlation lacking causality.

Our results, and particularly the fact that the stringency of the solvent accessibility term is specifically correlated with expression level (figure 4.7), are consistent with the recently proposed hypothesis that the underlying force behind the slower evolution of highly expressed proteins is the need to avoid misfolding. Misfolding may have particularly toxic effects in the case of abundant proteins. Aggregation -the association of several non-native protein molecules to form insoluble amorphous structures- is largely driven by hydrophobic forces (Hartl and Hayer-Hartl, 2009), which establishes a direct link with the solvent accessibility terms of the potentials. Exposure of particular combination of amino acids in the surface may thus induce the formation of aggregates (Bucciantini et al., 2002; Dobson, 2003). Conversely, non-favorable mutations in the hydrophobic core of the molecule would hinder the formation of a well packed core, producing partially unfolded ensembles that are prone to aggregation (reviewed in Rousseau et al., 2006). To further decompose the selective constraints involved, the data sets used here could be partitioned according to the solvent exposure of the sites, to analyze the posterior distribution of β_{solv} separately for sites located in the hydrophobic core and on the surface of the molecule.

The misfolding hypothesis mentioned above led Drummond and Wilke (2009) to propose that highly expressed proteins are more 'translationally robust', that is, more tolerant to amino acid change. Our results, on the other hand, point to a stronger purifying selection at the structural level, which translates into a lower tolerance to errors: a higher β implies that the fitness cost of a destabilizing change will be amplified in abundant proteins, in a way that is directly dependent on the energetic interactions of the molecule. The lower aggregation propensity of abundant proteins would then result from more stringent requirements on the accepted evolutionary changes conditional on the particular structural context, i.e. a higher sensitivity to errors.

4.5 Conclusion

Our results are exciting, but preliminary. For the particular problem of the underlying cause of the covariation of expression level and evolutionary rate, further evidence is required. The analysis presented here should be applied to a larger set of proteins to confirm the results observed. On a different direction, extensions to this approach should be envisioned to progressively consider additional factors. Some extensions can be easily introduced, such as parameters for describing codon usage bias (Yang and Nielsen, 2008), a measure also known to correlate with protein abundance. Some others, like the importance of protein-protein interactions, or the effect of folding kinetics, are more difficult to separate from the structural constraints modeled here. In any case, definite dissection of the different factors contributing to the strong selection on abundant proteins will require further work, and the modeling approach presented here may provide more direct means of testing alternative evolutionary hypothesis.

4.6 Supplementary figures and tables

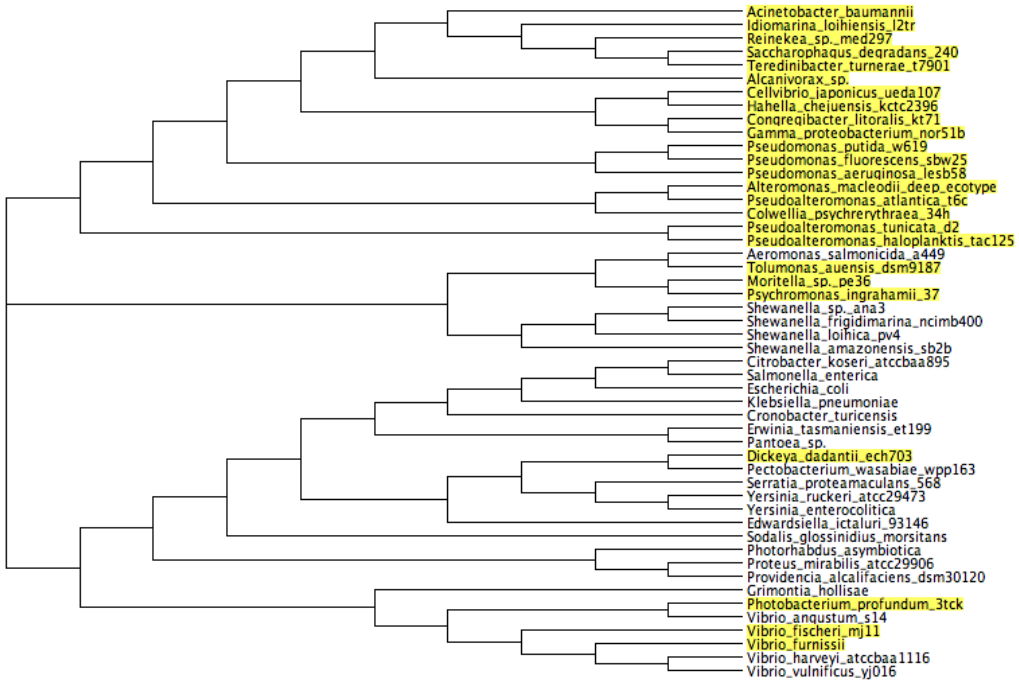


Figure 4.8: **Starting set of taxa used in the phylogenetic analysis.** Yellow colored species were removed from some of the datasets in order to reduce computation times (see Methods).

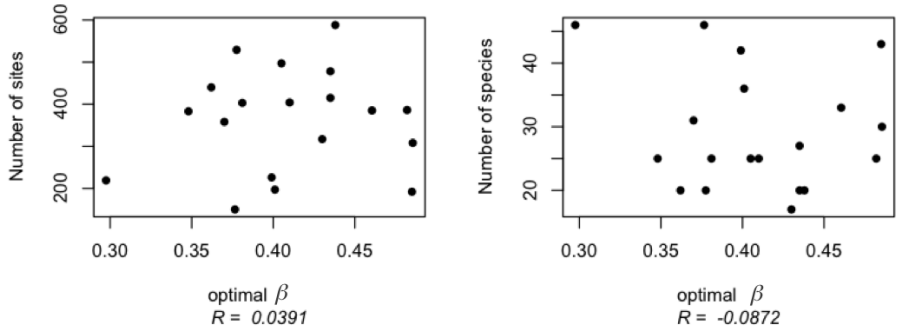


Figure 4.9: **Independence of optimal β and size of the dataset**, under MG-SC-NS model. Left panel: optimal β plotted against alignment length. Right panel: optimal β plotted against the number of species in the dataset. Correlation coefficients are shown at the bottom of each plot.

Table 4.III: **Optimal β.**

Gene	MG-S-SC		MG-NS-SC	
tufB	0.492	0.485	0.459	0.462
gapA	0.455	0.456	0.483	0.488
rpsE	0.361	0.366	0.377	0.376
serA	0.405	0.405	0.410	0.410
sodB	0.491	0.521	0.478	0.492
icdA	0.414	0.421	0.464	0.406
rpoA	0.388	0.391	0.397	0.401
fba	0.355	0.365	0.370	0.370
fabB	0.392	0.390	0.382	0.380
pgk	0.458	0.454	0.484	0.480
Average	0.42	0.43	0.43	0.43
sdhA	0.434	0.436	0.432	0.444
aceB	0.378	0.384	0.370	0.385
pepP	0.364	0.370	0.360	0.364
metB	0.358	0.362	0.344	0.352
atpG	0.315	0.316	0.301	0.294
nadB	0.404	0.410	0.434	0.436
acpD	0.392	0.418	0.384	0.418
asnB	0.396	0.396	0.402	0.408
wcaG	0.430	0.426	0.420	0.440
Average	0.39	0.39	0.38	0.39

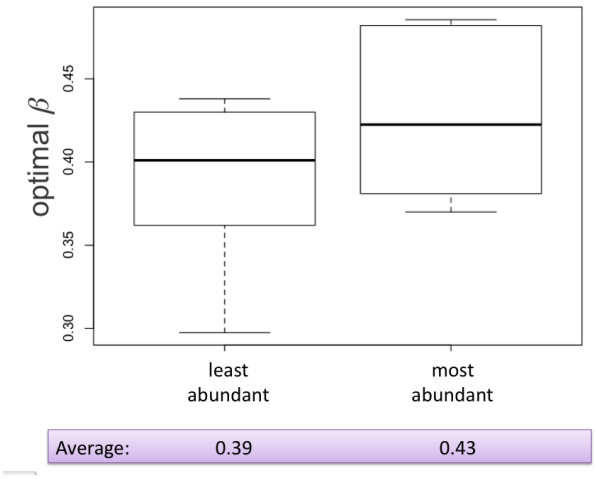


Figure 4.10: **Boxplots of optimal β for genes of high and low expression levels,** under MG-NS-SC model. *p-value* = 0.079.

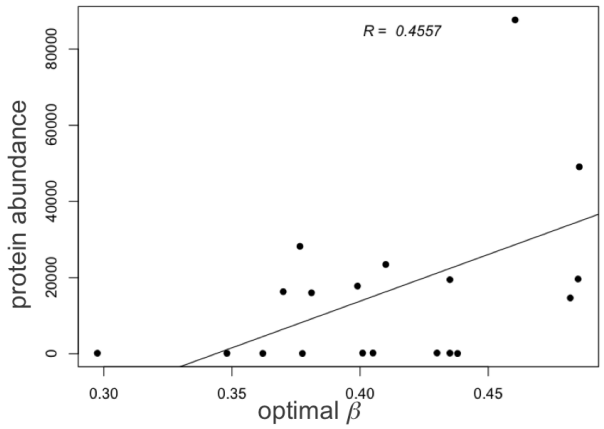


Figure 4.11: **Correlation of optimal β and protein abundance.** MG-NS-SC model was used.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

The main motivation of this work has been to incorporate detailed protein structure information into existing phylogenetic methods, using a unified model-based statistical framework to assess the relevance of this information. To what extent are the factors known to affect protein structure -in vitro, in isolation and controlled laboratory conditions- shaping the evolution of protein sequences? Can we disentangle selective constraints on protein structure from other selective forces? With these questions in mind, we developed a probabilistic framework to derive statistical potentials (Chapter 2 and Appendix 1) with detailed structural descriptions (Chapter 3), optimized for evolutionary studies. We incorporated the potentials into a structurally constrained model of sequence evolution and evaluated them in a Bayesian framework (Chapter 3). We then applied the inference framework to proteins of differing expression level, to better understand the evolutionary pressures that different proteins are subject to (Chapter 4), and the role that tertiary structure plays in this process.

Although a variety of available statistical potentials could have been used, the probabilistic formulation presented in chapter 2 provided us with reliable criteria to choose from a number of models on real proteins, refining the structural description so as to increase the amount of information included without overfitting the data or inflating computational times. It was general enough to allow for flexibility on the level of detail of the structural information included and on the type of terms that could be considered (not always directly related to the conformational free energy of the proteins). Furthermore, it provided us with control over the training databases, allowing us, for example, to correct for compositional effects of nucleotide sequences of the training proteins due to mutational pressures (Bonnard (2010) and chapter 4), increasing the fit of the model and the self-consistency of the evolutionary framework as a whole.

Previous to this work (Robinson et al., 2003; Rodrigue et al., 2005; Choi et al., 2007), only simple protein structure representations had been used. After refining substantially the structural descriptions, the fit of the models increases considerably. We find almost invariantly, however, that this explicit modeling approach improves the model fit with respect to simpler DNA models and existing SC models, but does not, on its own, outperform the best existing phenomenological alternatives. A first explanation for this result may be that the necessary simplifications made to derive the potentials make them insufficient to accurately model selective constraints related to the protein structure when applied to specific proteins. Although the potentials we developed include a great amount of structural information, they still keep the description at a coarse-grained level and are designed to capture general trends of amino acid propensities for average proteins, well represented in the learning database. Alternatively, the relatively mild performance of the SC models may come from the fact that they model selective constraints exclusively related to the protein structure, while in reality these are only a small fraction of all the selective constraints operating on sequences. If the critical interactions for maintaining a protein fold are relatively sparse, that is, affecting only a minority of sites (as suggested by the sequence logos of natural sequences displayed in chapters 2 and 3), model fit will be only slightly improved by the addition of structural specifications, no matter how accurate they are. The methodological tools we currently have do not allow us to distinguish between these two alternative hypotheses, and the true scenario is probably a combination of both.

In any case, as illustrated by the results presented in chapter 4, such an explicit modeling approach is attractive to study particular aspects of molecular evolution in a more direct way than with phenomenological alternatives. Several directions for future work that could be pursued will be detailed in the present chapter, both for improving the performance of the models as well as for applying them to biological problems.

5.1 Optimization procedure

In chapter 3 we focused on improving the functional form of the energy term, without attempting to change any aspect of the optimization itself. In Appendix 1, one major improvement is presented, where the convergence time is reduced by three orders of magnitude while obtaining essentially the same parameter values, by considering an alternative definition of the likelihood function. But there are other aspects of the optimization process that could benefit from further development.

Negative design

A sequence s designed for a target conformation c should not only be compatible with c, but also less compatible with competing folds (i.e. meeting *specificity* requirements). Performing searches in sequence space can yield sequences with lower energy than the starting sequence, but that are not a good alternative to it because they will fold into a different structure. A rigorous solution to this problem would thus involve a simultaneous search over the sequence-structure space, unfeasible but for small on-lattice proteins (Seno et al., 1996). Although we have shown that, relying on the approximation based on the random energy model (Shakhnovich and Gutin, 1993; Sun et al., 1995; Seno et al., 1998), it is possible to achieve specificity without explicitly considering competing conformational states (chapter 2), it is worth exploring if doing so improves the performance of the method.

The group of Nicolas Lartillot has started working in this direction (Bonnard, 2010), formalizing the problem to explicitly penalize non-native competing structures -called *decoys*- on the definition of the conditional probability of a sequence:

$$p(s \mid c) \propto e^{-[E(S,C)-\langle E(S)\rangle]} p(s) \tag{5.1}$$

where the expectation $\langle \cdot \rangle$ is taken over a pre-defined set of decoy conformations. The results obtained so far are promising, although further work is needed in order to define a proper set of decoys for this problem. In particular, by defining a set of decoys that represent the unfolded state and folding intermediates, we can begin to model constraints related to the mechanism of folding, which we are not taking into account so far. In our current formulation, only the folded state of proteins is modeled, by calculating the pseudo-energy change before and after a substitution, while the unfolded state is assumed to behave as a random polymer (by using the random energy model approximation). The exact nature of the unfolded state is an open question in structural biology, and though several advances have been made in recent years that could eventually be adapted to our purposes (Suárez and Jaramillo, 2009; Dill et al., 2008), this will certainly represent a non trivial task.

Conformational flexibility

The optimization procedure we developed considers one single native protein structure for each sequence. However, the native state is not a rigid object. It is most probably an ensemble of conformations, which in turn undergoes fluctuations during the evolutionary time scales we are encompassing. Thus, use of conformational ensembles should be an important component of modeling proteins in evolution. The coarse grained representation of the structure provides an indirect way of allowing flexibility, but given its importance for protein function, an explicit modeling of this feature would be desirable. We attempted to include it through the analysis of B-factors, but other implementations, directly on the optimization itself, can be envisioned.

Conformational diversity of the training proteins can be incorporated by generating ensembles from a single structure, for example by variation along normal modes or molecular dynamics simulations (reviewed in Friedland and Kortemme (2010)). Alternatively, it can be considered by integrating conformational diversity already present in the database, represented by different states of the same protein, or by homologous

structures. These two alternatives are in some cases equivalent. It has been shown a number of years ago that conformations sampled in a molecular dynamics trajectory of myoglobin are similar to the diversity present in crystal structures of its family (Elber and Karplus, 1987). More recently, Vendruscolo and colleagues observed that ensembles of crystallographic structures of proteins with high sequence identity can be representative of NMR dynamics measurements (Best et al., 2006), and considerable similarity has been shown between a modeled conformational ensemble and an ensemble of structures of members of the protein's natural family (Friedland et al., 2009).

In our context, this structural ensemble could be used in the training database instead of the single native structure (for ways to represent these ensembles, see figure 5.1 and Friedland and Kortemme (2010). One possible way to achieve this is to use, instead of the native conformation associated to each sequence, for each structural descriptor a pondered mean of several conformations. Panjkovich et al. (2008), in the same spirit, derived knowledge-based potentials using a single experimental structure and all three dimensional models of its homologous sequences.

5.2 Structural description

From the initial formulation of the potential energy based solely on binary contact interactions, the structural description included in the SC models has been progressively ameliorated by the inclusion of terms related to distance-based interactions, solvent accessibility, torsion angles and flexibility of the residues. Each one of these terms was in turn carefully refined so as to maximize predictions. Further improvements could be made in this direction. For example, it is known that the allowed Ramachandran regions for a site are affected by the identity and conformation of neighboring sites (Zaman et al., 2003). This could be incorporated in the potentials by including terms related to interactions in torsion angles among successive positions along the chain (Betancourt and Skolnick, 2004). Regarding the treatment of pairwise interactions, sequence sep-

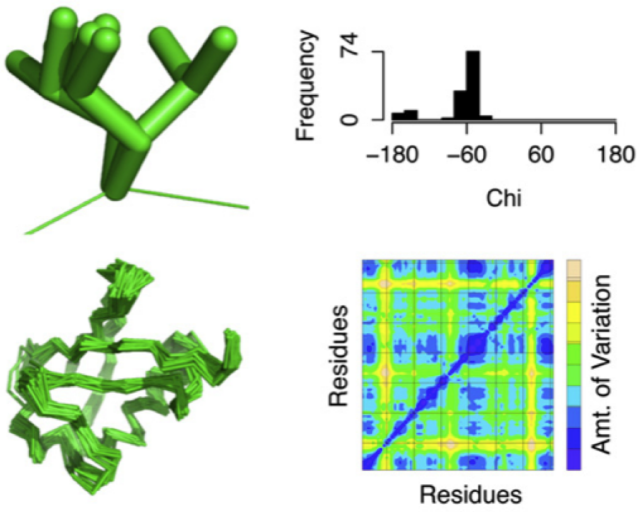


Figure 5.1: **Several ways of describing conformational ensembles.** Multiple side chain conformations on a fixed backbone (top left) can be represented by a histogram of dihedral angles (top right), and multiple backbone conformations (bottom left) can be represented by an average C-alpha distance difference matrix (bottom right). A C-alpha distance difference matrix describes the amount of variation in the distance between pairs of residues in an ensemble. Figure adapted from Friedland and Kortemme (2010).

aration ranges can be considered (Sippl, 1993), distinguishing among interactions of residues close in sequence number (which have a higher probability of ocurring) and long-range interactions. Also, sidechain-backbone interactions (Buchete et al., 2004) can be specifically modeled. Furthermore, the method can be extended to consider interactions involving more than two residues at a time. Up-to four body potential energies have been developed with a similar formulation to ours (Krishnamoorthy and Tropsha, 2003), showing improvements when discriminating native from misfolded structures.

All these extensions can be relatively easily implemented and tested using the methodological framework established in chapter 2 for model assessment and comparison. It is not clear however, given the results obtained within the phylogenetic context (chapter 3 and 4), to what extent these relatively minor refinements of the current formulation will

produce significant improvements on the fit of SC models, to the point of making them competitive against current phenomenological alternatives. We are not too optimistic on this point.

On the other hand, the results we obtained (figures 3.2 and 3.3) show that the most important steps in model fit are brought about by the distance-based potentials and, among all the versions tested, by the one including the most information about the position of the side chain. This suggests that reducing the coarse-grained nature of the potentials may indeed make a difference. In this direction, further information on side chain conformations may be included, for example by considering terms related to the dihedral angles around the bonds of the side chain atoms (χ^1, χ^2, etc.), in the same way as the modeling of main-chain torsion angles in the current formulation. In any case, the maximum likelihood framework proposed is very general, and is not restricted to coarse-grained descriptions of proteins. An analogous statistical potential, formulated at the atomic level, may be worth developing.

5.3 Applications and model extensions

Efforts to incorporate phenotype into evolutionary studies are at an early stage. For site-interdependent models, most of the work so far has focused on developing the computational tools to perform phylogenetic inference and model comparison dealing with site dependences. The properties of the models though, especially when applied to large and heterogeneous data sets, are still virtually unexplored.

We have started to pursue this line of research with the pilot study presented in chapter 4. Clearly, this study needs to be extended to a more representative number of proteins to confirm the tendencies observed. Regardless, it is a good illustration of how SC models can already be used to quantitatively assess variations of evolutionary pressures related to protein structure. Besides being extended to a larger scale, the analysis should also be performed for eukaryotic organisms, since their mechanisms of protein synthesis,

folding and prevention of misfolding are quite different from those of bacteria.

No longer restricting the analysis to proteins encoded within the same genome, the variability of selective pressures associated with protein structure can also be analyzed in relation to other variables, besides protein abundance. It will be interesting to see to what extent a mechanistic model of evolution articulating the combined effects of mutational pressure and natural selection allows us to make a connection between the substitution process and ecological and physiological conditions that can differ substantially between taxonomic groups. A trivial example would be the effect of temperature: thermophiles should display a more stringent selection on protein structure than mesophiles. But other more subtle connections can be established. For example, it is known that selection intensity is proportional to effective population size, and thus, it will be interesting to see how this impacts structural constraints. A first preliminary analysis can be performed in the same spirit as the one presented in chapter 4, contrasting posterior distributions of parameters describing selection (β_x) in groups differing in population size (mammals, vertebrates, animals, yeast, bacteria). A more sophisticated approach can be devised in a later stage to take into account the variation of population size over time, and its relation to selection on protein structure.

In a different perspective, extensions of the SC models can now be constructed to progressively incorporate additional selective constraints. In general, this modeling approach may be extended to include any phenotypic trait that can be predicted from genotype. Mapping genotype to phenotype and phenotype to fitness is, unfortunately, not a trivial task, although there are some cases that could be easily implemented. Selection for translation efficiency and accuracy, for example, can be modeled using parameters describing codon usage bias (Yang and Nielsen, 2008; Rodrigue et al., 2008a). The impact of RNA secondary structure on sequence change can be modeled in a similar way as the impact of protein tertiary structure (Yu and Thorne, 2006). Immunogenic properties (which are subject to negative selection in the case of pathogens) can be predicted from sequence (Moutaftsi et al., 2006) and so incorporated into the model, also with a

similar formulation as the one we used. Folding constraints, in turn, could eventually be included by modeling the unfolded state and folding intermediates in the statistical potential, as discussed above.

Much progress is being made in separately modeling diverse aspects of sequence evolution, and in incorporating a part of the wealth of biological knowledge available. I presented here our work in this context, restricted to structural properties of proteins. The challenge now is to combine all the separate pieces into a unified and coherent model, to improve our understanding of how each different factor is contributing to the evolution of protein sequences.

BIBLIOGRAPHY

Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1996. Improved design of stable and fast-folding model proteins. *Fold Des* 1:221–230.

Adachi, J. 1996. Model of amino acid substitution in proteins encoded by mitochondrial dna. *J Mol Evol* 42:459–468.

Adachi, J., P.J. Waddell, W. Martin, and M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast dna. *J Mol Evol* 50:348–358.

Akashi, H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Devel* 11:660–666.

Anisimova, M., and C. Kosiol. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol* 26:255–71.

Aris-Brosou, S. 2005. Determinants of Adaptive Evolution at the Molecular Level: the Extended Complexity Hypothesis. *Mol Biol Evol* 22:200–209.

Artymiuk, P. J., C. C. Blake, D. E. Grace, S. J. Oatley, D. C. Phillips, and M. J. Sternberg. 1979. Crystallographic studies of the dynamic properties of lysozyme. *Nature* 280:563–8.

Auger, S., A. Danchin, and I. Martin-Verstraete. 2002. Global expression profile of bacillus subtilis grown in the presence of sulfate or methionine. *J Bacteriol* 184:5179–5186.

Banavar, J.R., M. Cieplak, A. Maritan, G. Nadig, F. Seno, and S. Vishveshwara. 1998. Structure-based design of model proteins. *Proteins* 32:80–87.

Bastolla, U., J. Farwer, E. W. Knapp, and M. Vendruscolo. 2001. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins* 44:79–96.

Bastolla, U., M. Porto, H. E. Roman, and M. Vendruscolo. 2002. Lack of self-averaging in neutral evolution of proteins. *Phys Rev Lett* 89.

Bastolla, U., M. Porto, H. E. Roman, and M. Vendruscolo. 2003. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *J Mol Evol* 56:243–254.

Bastolla, U., M. Vendruscolo, and E. W. Knapp. 2000. A statistical mechanical method to optimize energy functions for protein folding. *P Natl Acad Sci USA* 97:3977–81.

Ben-Naim, A. 1997. Statistical potentials extracted from protein structures: Are these meaningful potentials? *J Chem Phys* 107:3698–3706.

Bergman, N. H., Erica C. Anderson, E. E. Swenson, B. K. Janes, N. Fisher, M. M. Niemeyer, A. D. Miyoshi, and P. C. Hanna. 2007. Transcriptional profiling of bacillus anthracis during infection of host macrophages. *Infect Immun* 75:3434–3444.

Best, R.B., K. Lindorff-Larsen, M.A. DePristo, and M. Vendruscolo. 2006. Relation between native ensembles and experimental structures of proteins. *P Natl Acad Sci USA* 103:10901.

Betancourt, A. J., and D. C. Presgraves. 2002. Linkage limits the power of natural selection in Drosophila. *P Natl Acad Sci USA* 99:13616–13620.

Betancourt, M. R., and J. Skolnick. 2004. Local propensities and statistical potentials of backbone dihedral angles in proteins. *J Mol Biol* 342:635–49.

Blaisdell, B.E. 1985. A method for estimating from two aligned present day dna sequences their ancestral composition and subsequent rates of composition and subse-

quent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site. *J Mol Biol* 22:69–81.

Blanquart, S., and N. Lartillot. 2006. A bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol Biol Evol* 23:2058–2071.

Boas, F. E., and P. B. Harbury. 2007. Potential energy functions for protein design. *Curr Opin Struc Biol* 17:199–204.

Bolon, D. N., R. A. Grant, T. A. Baker, and R. T. Sauer. 2005. Specificity versus stability in computational protein design. *P Natl Acad Sci USA* 102:12724–9.

Bonnard, C. 2010. Optimisation de potentiels statistiques pour un modèle d'évolution soumis à des contraintes structurales. Phd dissertation, Université Montpellier II.

Bonnard, C., C. L. Kleinman, N. Rodrigue, and N. Lartillot. 2009. Fast optimization of statistical potentials for structurally constrained phylogenetic models. *BMC Evol Biol* 9:227.

Boussau, B., and M. Gouy. 2006. Efficient Likelihood Computations with Nonreversible Models of Evolution. *Syst Biol* 55:756–768.

Bradley, R. K., A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, and L. Pachter. 2009. Fast statistical alignment. *PLoS Comput Biol* 5:e1000392.

Brooks, B. R., C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, and M. York, D. M. andKarplus. 2009. Charmm: The biomolecular simulation program. *J Comput Chem* 30:1545–1614.

Brooks, B.R., R.E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217.

Bruno, W.J. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol Biol Evol* 13:1368–1374.

Bucciantini, M., E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson, and M. Stefani. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416:507–11.

Buchete, N. V., J. E. Straub, and D. Thirumalai. 2004. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci* 13:862–74.

Bustamante, C. D., J. P. Townsend, and D. L. Hartl. 2000. Solvent accessibility and purifying selection within proteins of escherichia coli and salmonella enterica. *Mol Biol Evol* 17:301–8.

Cherry, Joshua L. 2010. Highly Expressed and Slowly Evolving Proteins Share Compositional Properties with Thermophilic Proteins. *Mol Biol Evol* 27:735–741.

Chiu, T. L., and R. A. Goldstein. 1998. Optimizing potentials for the inverse protein folding problem. *Protein Eng* 11:749–52.

Choi, S. C., A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol* 24:1769–82.

Choi, S. S., E. J. Vallender, and B. T. Lahn. 2006. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. *Mol Biol Evol* 23:2131–3.

Christensen, O. F., A. Hobolth, and J. L. Jensen. 2005. Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *J Comput Biol* 12:1166–1182.

Conant, G. C., and P. F. Stadler. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol Biol Evol* 26:1155–61.

Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, Jr. Merz, K. M., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117:5179–97.

Dahiyat, B. I., C. A. Sarisky, and S. L. Mayo. 1997. De novo protein design: towards fully automated sequence selection. *J Mol Biol* 273:789–796.

Dayhoff, M., R. Schwartz, and B. Orcutt. 1978. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, ed. M. Dayhoff, 345–352. National Biomedical Research Foundation, Washington, D.C.

Dean, A.M., C. Neuhauser, E. Grenier, and G.B. Golding. 2002. The pattern of amino acid replacements in α/β-barrels. *Mol Biol Evol* 19:1846–1864.

Delport, W., K. Scheffler, and C. Seoighe. 2009. Models of coding sequence evolution. *Brief Bioinform* 10:97–109.

Dempster, A., N Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39:1–38.

Deutsch, J. M., and T. Kurosky. 1996. New algorithm for protein design. *Phys Rev Lett* 76:323–326.

Dill, Ken A., S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. 2008. The protein folding problem. *Ann Rev Biophys* 37:289–316.

Dimmic, M. W., D. P. Mindell, and R. A. Goldstein. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac Symp Biocomput* 18–29.

Dobson, C. M. 2003. Protein folding and misfolding. *Nature* 426:884–90.

Drexler, K. E. 1981. Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *P Natl Acad Sci USA* 78:5275–5278.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. *P Natl Acad Sci USA* 102:14338–14343.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23:327–37.

Drummond, D. A., and C. O. Wilke. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–52.

Drummond, D.A., and C.O. Wilke. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* 10:715–724.

Dunbrack, R. L., and M. Karplus. 1993. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J Mol Biol* 230:543 – 574.

Duret, L., and D. Mouchiroud. 2000. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Mol Biol Evol* 17:68–70.

Edgar, R. C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–7.

Elber, R., and M. Karplus. 1987. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science* 235:318–318.

Ellegren, H., N. G. C. Smith, and M. T. Webster. 2003. Mutation rate variation in the mammalian genome. *Curr Opin Genet Devel* 13:562 – 568.

Felsenstein, J. 1981. Evolutionary trees from dna sequences: A maximum likelihood approach. *J Mol Biol* 17:368–376.

Felsenstein, J. 2004. *Inferring phylogenies / joseph felsenstein*. Sunderland, Mass. : Sinauer Associates.

Flores, T. P., C. A. Orengo, D. S. Moss, and J. M. Thornton. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci* 2:1811–26.

Foster, Peter G. 2004. Modeling Compositional Heterogeneity. *Syst Biol* 53:485–495.

Franzosa, Eric A., and Yu Xia. 2009. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Mol Biol Evol* 26:2387–2395.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary Rate in the Protein Interaction Network. *Science* 296:750–752.

Frauenfelder, H., G. A. Petsko, and D. Tsernoglou. 1979. Temperature-dependent x-ray diffraction as a probe of protein structural dynamics. *Nature* 280:558–63.

Friedland, G.D., and T. Kortemme. 2010. Designing ensembles in conformational and sequence space to characterize and engineer proteins. *Curr Opin Struc Biol* 20:377–384.

Friedland, G.D., N.A. Lakomek, C. Griesinger, J. Meiler, and T. Kortemme. 2009. A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comput Biol* 5:e1000393.

Galtier, N, and M Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol* 15:871–879.

Galtier, Nicolas. 2001. Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. *Mol Biol Evol* 18:866–873.

Gatchell, D.W., S. Dennis, and S. Vajda. 2000. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* 41:518–534.

Gelman, A. 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Stat Sci* 13:163–185.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A., X.L. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Satistica Sinica* 6:733–760.

Gerstein, M., and N. Echols. 2004. Exploring the range of protein flexibility, from a structural proteomics perspective. *Curr Opin Chem Biol* 8:14–19.

Gilis, D., and M. Rooman. 1997. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J Mol Biol* 272:276–90.

Gilis, D., and M. Rooman. 2001. Identification and ab initio simulations of early folding units in proteins. *Proteins* 42:164–76.

Glaser, F., Y. Rosenberg, A. Kessel, T. Pupko, and N. Ben-Tal. 2005. The consurf-hssp database: the mapping of evolutionary conservation among homologs onto pdb structures. *Proteins* 58:610–7.

Godzik, A., A. Kolinski, and J. Skolnick. 1995. Are proteins ideal mixtures of amino acids? analysis of energy parameter sets. *Protein Sci* 4:2107–2117.

Gojobori, T., K. Ishii, and M. Nei. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J Mol Evol* 18:414–423.

Goldman, N., J. L. Thorne, and D. T. Jones. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J Mol Biol* 263:196–208.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445–58.

Goldman, N, and Z Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725–736.

Gong, S., C. L. Worth, G. R. Bickerton, S. Lee, D. Tanramluk, and T. L. Blundell. 2009. Structural and functional restraints in the evolution of protein families and superfamilies. *Biochem Soc Trans* 37:727–33.

Gordon, D. B., S. A. Marshall, and S. L. Mayo. 1999. Energy functions for protein design. *Curr Opin Struc Biol* 9:509 – 513.

Gowri-Shankar, Vivek, and Magnus Rattray. 2007. A Reversible Jump Method for Bayesian Phylogenetic Inference with a Nonhomogeneous Substitution Model. *Mol Biol Evol* 24:1286–1299.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.

Guvench, O., and A.D. Jr. MacKerell. 2008. Comparison of protein force fields for molecular dynamics simulations. In *Molecular modeling of proteins*, ed. Andreas Kukol, 63–88. Humana Press.

Halpern, AL, and WJ Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site- specific residue frequencies. *Mol Biol Evol* 15:910–917.

Haney, P.J., J.H. Badger, G.L. Buldak, C.I. Reich, C.R. Woese, and G.J. Olsen. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic methanococcus species. *P Natl Acad Sci USA* 96:3578–3583.

Hartl, F.U., and M. Hayer-Hartl. 2009. Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol* 16:574–581.

Hasegawa, M., H. Kishino, and T.A. Yano. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. *J Mol Evol* 22:160–174.

Hellinga, H. W., and F. M. Richards. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *P Natl Acad Sci USA* 91:5803–5807.

Hendlich, M., P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. 1990. Identification of native protein folds amongst a large number of incorrect models. the calculation of low energy conformations from potentials of mean force. *J Mol Biol* 216:167–180.

Herbeck, JT, and DP Wall. 2005. Converging on a general model of protein evolution. *Trends Biotechnol* 23:485–487.

Hirsh, A.E., and H.B. Fraser. 2001. Protein dispensability and rate of evolution. *Nature* 411:1040–1049.

Hoeflich, K. P., and M. Ikura. 2002. Calmodulin in action: diversity in target recognition and activation mechanisms. *Cell* 108:739–42.

Hu, H., M. Elstner, and J. Hermans. 2003. Comparison of a qm/mm force field and molecular mechanics force fields in simulations of alanine and glycine dipeptides

(ace-ala-nme and ace-gly-nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins* 50:451–463.

Hubbard, S.J., and J. M. Thornton. 1993. Naccess. Depart. of Biochem. and Molec. Biol., University College, London.

Huelsenbeck, J. P., S. Jain, S. W. Frost, and S. L. Pond. 2006. A dirichlet process model for detecting positive selection in protein-coding dna sequences. *P Natl Acad Sci USA* 103:6263–8.

Huelsenbeck, John P. 2002. Testing a Covariotide Model of DNA Substitution. *Mol Biol Evol* 19:698–707.

Huelsenbeck, John P., Bret Larget, Richard E. Miller, and Fredrik Ronquist. 2002. Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. *Syst Biol* 51:673–688.

Huse, M., and J. Kuriyan. 2002. The conformational plasticity of protein kinases. *Cell* 109:275–282.

Hwang, Dick G., and Phil Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *P Natl Acad Sci USA* 101:13994–14001.

Jaramillo, A., L. Wernisch, S. Héry, and S. J. Wodak. 2002. Folding free energy function selects native-like protein sequences in the core but not on the surface. *P Natl Acad Sci USA* 99:13554–13559.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. *Math Proc Cambridge* 31:203–222.

Jensen, J. L., and A. K. Pedersen. 2000. Probabilistic models of dna sequence evolution with context dependent rates of substitution. *Adv Appl Prob* 32:499–517.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. Treefinder: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* 4:18.

Jones, D. T. 1997. Successful ab initio prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* Suppl 1:185–91.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992a. A new approach to protein fold recognition. *Nature* 358:86–9.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992b. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–82.

Jorgensen, W. L., and J. Tirado-Rives. 1988. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 110:1657–66.

Jukes, T.H., and C.R. Cantor. 1969. Evolution of protein molecules. In *Mammalian protein metabolism*, ed. Munro H.N., 21–132. Academic Press, New York, NY.

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

Karplus, M., and G.A. Petsko. 1990. Molecular dynamics simulations in biology. *Nature* 347:631–639.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *J Am Stat Assoc* 90:773–795.

Kim, S.H., and S.V. Yi. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. *Genetica* 131:151–156.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120.

King Jordan, I., I.B. Rogozin, Y.I. Wolf, and E.V. Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* 12:962–968.

Kleinman, C. L., N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot. 2006. A maximum likelihood framework for protein design. *BMC Bioinformatics* 7:326.

Kleinman, C. L., N. Rodrigue, N. Lartillot, and H. Philippe. 2010. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol* 27:1546–1560.

Kocher, Jean-Pierre A., Marianne J. Rooman, and Shoshana J. Wodak. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol* 235:1598–1613.

Koehl, P., and M. Levitt. 1999. De novo protein design. I. in search of stability and specificity. *J Mol Biol* 293:1161–1181.

Kono, H., and J. G. Saven. 2001. Statistical theory for protein combinatorial libraries. packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol* 306:607–628.

Koonin, E. V., and Y. I. Wolf. 2006. Evolutionary systems biology: links between gene evolution and function. *Curr Opin Biotechnol* 17:481–487.

Koshi, J. M., and R. A. Goldstein. 1995. Context-dependent optimal substitution matrices. *Protein Eng* 8:641–5.

Krishnamoorthy, B., and A. Tropsha. 2003. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics* 19:1540–1548.

Krylov, DM, YI Wolf, IB Rogozin, and EV Koonin. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* 13:2229–2235.

Kuhlman, B., and D. Baker. 2000. Native protein sequences are close to optimal for their structures. *P Natl Acad Sci USA* 97:10383–8.

Kurosky, T., and J. M. Deutsch. 1995. Design of copolymeric material. *J Phys A: Math Gen* 27:L387–L393.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal w and clustal x version 2.0. *Bioinformatics* 23:2947–8.

Larson, S. M., J. L. England, J. R. Desjarlais, and V. S. Pande. 2002. Thoroughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci* 11:2084–2813.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* 7:S4.

Lartillot, N., and H. Philippe. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095–109.

Lartillot, N., and H. Philippe. 2006. Computing bayes factors using thermodynamic integration. *Syst Biol* 55:195–207.

Laskowski, R. A. 2009. Pdbsum new things. *Nucleic Acids Res* 37:D355–9.

Laskowski, R. A., V. V. Chistyakov, and Thornton J. M. 2005. Pdbsum more: new summaries and analyses of the known 3d structures of proteins and nucleic acids. *Nucleic Acids Res* 33:D266–D268.

Laskowski, R. A., M. W. Macarthur, D. S. Moss, and J. M. Thornton. 1993. Procheck - a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26:283–291.

Laskowski, R. A., J. A. Rullmannn, M. W. MacArthur, R. Kaptein, and J. M. Thornton. 1996. Aqua and procheck-nmr: programs for checking the quality of protein structures solved by nmr. *J Biomol NMR* 8:477–86.

Lazaridis, T., and M. Karplus. 2000. Effective energy functions for protein structure prediction. *Curr Opin Struc Biol* 10:139–45.

Lee, B., and M. Richards. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol* 55:379–400.

Lemos, B, BR Bettencourt, CD Meiklejohn, and DL Hartl. 2005. Evolution of proteins and gene expression levels are coupled in drosophila and are independently associated with mrna abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol* 22:1345–1354.

Lercher, M. H., and L. D. Hurst. 2002. Human snp variability and mutation rate are higher in regions of high recombination. *Trends Genet* 18:337 – 340.

Lercher, Martin J., Elizabeth J. B. Williams, and Laurence D. Hurst. 2001. Local Similarity in Evolutionary Rates Extends over Whole Chromosomes in Human-Rodent and Mouse-Rat Comparisons: Implications for Understanding the Mechanistic Basis of the Male Mutation Bias. *Mol Biol Evol* 18:2032–2039.

Lio, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. *Genome Res* 8:1233–44.

Lio, P., N. Goldman, J. L. Thorne, and D. T. Jones. 1998. Passml: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* 14:726–33.

Lu, P., C. Vogel, R. Wang, X. Xao, and E.M. Marcotte. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech* 25:117–124.

Madhusudhan, M. S., M. A. Marti-Renom, N. Eswar, B. John, U. Pieper, R. Karchin, M.i.n.-.y.i. Shen, and A. Sali. 2005. Comparative protein structure modeling. In *Proteomics protocols handbook*, ed. J.M. Walker, 831–860. Humana Press Inc., Totowa, NJ.

Maguid, S., S. Fernandez-Alberti, G. Parisi, and J. Echave. 2006. Evolutionary conservation of protein backbone flexibility. *J Mol Evol* 63:448–57.

Maiorov, V., and G. Crippen. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol* 227:876–888.

Margulies, E. H., and E. Birney. 2008. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet* 9:303–313.

Meller, J., and R. Elber. 2001. Linear optimization and a double statistical filter for protein threading protocols. *Proteins* 45:241–261.

Melo, F., R. Sanchez, and A. Sali. 2002. Statistical potentials for fold assessment. *Protein Sci* 11:430–48.

Micheletti, C., F. Seno, A. Maritan, and J. Banavar. 1998. Design of proteins with hydrophobic and polar amino acids. *Proteins* 32:80–87.

Mirny, L. A., and E. I. Shakhnovich. 1996. How to derive a protein folding potential? a new approach to an old problem. *J Mol Biol* 264:1164–1179.

Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18:534–552.

Miyazawa, S., and R. L. Jernigan. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–44.

Moutaftsi, M., B. Peters, V. Pasquetto, D.C. Tscharke, J. Sidney, H.H. Bui, H. Grey, and A. Sette. 2006. A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. *Nat Biotech* 24:817–819.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11:715–24.

Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat* 9:249–265.

Ogata, Y. 1989. A Monte Carlo method for high dimensional integration. *Numerische Mathematik* 55:137–157.

Olsen, G. J. 1987. Earliest Phylogenetic Branchings: Comparing rRNA-based Evolutionary Trees Inferred with Various Techniques. *Cold Spring Harb Sym* 52:825–837.

Overington, J., M. S. Johnson, A. Sali, and T. L. Blundell. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc Biol Sci* 241:132–45.

Pabo, C. 1983. Molecular technology: designing proteins and peptides. *Nature* 301:200.

Pal, C., B. Papp, and M. J. Lercher. 2006. An integrated view of protein evolution. *Nat Rev Genet* 7:337–48.

Pal, Csaba, Balazs Papp, and Laurence D. Hurst. 2001. Does the Recombination Rate Affect the Efficiency of Purifying Selection? The Yeast Genome Provides a Partial Answer. *Mol Biol Evol* 18:2323–2326.

Panjkovich, A., F. Melo, and M. Marti-Renom. 2008. Evolutionary potentials: structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs. *Genome Biology* 9:R68.

Parisi, G., and J. Echave. 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol* 18:750–6.

Park, S., X. Yang, and J. G. Saven. 2004. Advances in computational protein design. *Curr Opin Struc Biol* 14:487–494.

Pedersen, A. M., and J. L. Jensen. 2001. A dependent-rates model and an mcmc-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol Biol Evol* 18:763–76.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. *Annu Rev Ecol Evol S* 36:541–562.

Plotkin, J.B., and H.B. Fraser. 2007. Assessing the determinants of evolutionary rates in the presence of noise. *Mol Biol Evol* 24:1113.

Pond, S. K., and S. V. Muse. 2005. Site-to-Site Variation of Synonymous Substitution Rates. *Mol Biol Evol* 22:2375–2385.

Ponder, J. W., and D. A. Case. 2003. Force fields for protein simulations. *Adv Protein Chem* 66:27–85.

Ponders, J. W., and F. M. Richards. 1987. Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* 193:775–791.

Poole, A. M., and R. Ranganathan. 2006. Knowledge-based potentials in protein design. *Curr Opin Struc Biol* 16:508–513.

Potapov, V., M. Cohen, and G. Schreiber. 2009. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel* 22:553–560.

Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–9.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol* 20:1692–704.

Rodrigue, N., C. L. Kleinman, H. Philippe, and N. Lartillot. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol* 26:1663–76.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–17.

Rodrigue, N., N. Lartillot, and H. Philippe. 2008a. Bayesian comparisons of codon substitution models. *Genetics* 180:1579–91.

Rodrigue, N., and H. Philippe. 2010. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet* 26:248 – 252.

Rodrigue, N., H. Philippe, and N. Lartillot. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol* 23:1762–75.

Rodrigue, N., H. Philippe, and N. Lartillot. 2007. Exploring fast computational strategies for probabilistic phylogenetic analysis. *Syst Biol* 56:711–26.

Rodrigue, N., H. Philippe, and N. Lartillot. 2008b. Uniformization for sampling realizations of markov processes: applications to bayesian implementations of codon substitution models. *Bioinformatics* 24:56–62.

Rodrigue, N., H. Philippe, and N. Lartillot. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *P Natl Acad Sci USA* 107:4629–4634.

Rossi, A., A. Maritan, and C. Micheletti. 2000. A novel iterative strategy for protein design. *J Chem Phys* 112:2050–2055.

Rossi, A., C. Micheletti, F. Seno, and A. Maritan. 2001. A self-consistent knowledge-based approach to protein design. *Biophys J* 80:480–490.

Rousseau, F., J. Schymkowitz, and L. Serrano. 2006. Protein aggregation and amyloidosis: confusion of the kinds? *Curr Opin Struc Biol* 16:118–126.

Roux, B., and T. Simonson. 1999. Implicit solvent models. *Biophys Chem* 78:1–20.

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Stat* 4:1151–1172.

Russell, R. B., M. A. Saqi, R. A. Sayle, P. A. Bates, and M. J. Sternberg. 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J Mol Biol* 269:423–39.

Rykunov, D., and A. Fiser. 2010. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics* 128.

Santos, J., C. Marino-Buslje, C. L. Kleinman, M R. Ermácora, and J.M. Delfino. 2007. Consolidation of the thioredoxin fold by peptide recognition: Interaction between e. coli thioredoxin fragments 1?93 and 94?108. *Biochemistry* 46:5148–5159.

Schlessinger, A., and B. Rost. 2005. Protein flexibility and rigidity predicted from sequence. *Proteins* 61:115–26.

Schneider, T. D., and R. M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acid Res* 18:6097–6100.

Schramm, G., M. Zapatka, R. Eils, and R. Konig. 2007. Using gene expression data and network topology to detect substantial pathways, clusters and switches during oxygen deprivation of escherichia coli. *BMC Bioinformatics* 8:149.

Seno, F., C. Micheletti, A. Maritan, and J. R. Banavar. 1998. Variational approach to protein design and extraction of interaction potentials. *Phys Rev Lett* 81:2172–2175.

Seno, F., M. Vendruscolo, A. Maritan, and J. R. Banavar. 1996. Optimal protein design procedures. *Phys Rev Lett* 77:1901–1904.

Shakhnovich, B. E., E. Deeds, C. Delisi, and E. Shakhnovich. 2005. Protein structure and evolutionary history determine sequence space topology. *Genome Res* 15:385–92.

Shakhnovich, E. I., and A. M. Gutin. 1993. Engineering of stable and fast-folding sequences of model proteins. *P Natl Acad Sci USA* 90:7195–9.

Siepel, A., and D. Haussler. 2004. Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. *Mol Biol Evol* 21:468–488.

Sippl, M. J. 1990. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213:859–883.

Sippl, M. J. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *J Comput Aided Mol Des* 7:473–501.

Skolnick, J., L. Jaroszewski, A. Kolinski, and A. Godzik. 1997. Derivation and testing of pair potentials for protein folding. when is the quasi-chemical approximation correct? *Protein Sci* 6:676–688.

Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. 2005. Evolutionary information for specifying a protein fold. *Nature* 437:512–8.

Solis, A. D., and S. Rackovsy. 2006. Improvement of statistical potentials and threading score functions using information maximization. *Proteins* 62:892–908.

Sternberg, M. J., D. E. Grace, and D. C. Phillips. 1979. Dynamic information from protein crystallography. an analysis of temperature factors from refinement of the hen egg-white lysozyme structure. *J Mol Biol* 130:231–52.

Suárez, M., and A. Jaramillo. 2009. Challenges in the computational design of proteins. *Journal of The Royal Society Interface* 6:S477–S491.

Subramanian, S., and S. Kumar. 2004. Gene Expression Intensity Shapes Evolutionary Rates of the Proteins Encoded by the Vertebrate Genome. *Genetics* 168:373–381.

Suel, G. M., S. W. Lockless, M. A. Wall, and R. Ranganathan. 2002. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10:59–69.

Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. *Annu Rev Ecol Syst* 36:445–466.

Sun, S., R. Brem, H. S. Chan, and K. A. Dill. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng* 8:1205–13.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hills. 1996. Phylogenetic inference. In *Molecular systematics*, ed. J.M. Walker, 407–514. Sinauer Associates, Sunderland, Massachusetts.

Takahata, Naoyuki, and Motoo Kimura. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics* 98:641–657.

Tanaka, S., and H. A. Scheraga. 1976. Medium- and long-range interaction parameters between amino acids for preditcing three dimensional structures of proteins. *Macromolecules* 9:945–950.

Tartaglia, G. C., S. Pechmann, C. M. Dobson, and M. Vendruscolo. 2007. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem Sci* 32:204 – 206.

Tartaglia, G. C., S. Pechmann, C. M. Dobson, and M. Vendruscolo. 2009. A relationship between mrna expression levels and protein solubility in e. coli. *J Mol Biol* 388:381–389.

Taverna, D. M., and R. A. Goldstein. 2002a. Why are proteins marginally stable? *Proteins* 46:105–9.

Taverna, D. M., and R. A. Goldstein. 2002b. Why are proteins so robust to site mutations? *J Mol Biol* 315:479–84.

Thomas, P. D., and K. A. Dill. 1996a. An iterative method for extracting energy-like quantities from protein structures. *P Natl Acad Sci USA* 93:11628–11633.

Thomas, P. D., and K. A. Dill. 1996b. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol* 257:457–469.

Thorne, J. L., S. C. Choi, J. Yu, P. G. Higgs, and H. Kishino. 2007. Population genetics without intraspecific data. *Mol Biol Evol* 24:1667–77.

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. *Mol Biol Evol* 13:666–73.

Tiana, G., M. Colombo, D. Provasi, and R. A. Broglia. 2004. Deriving amino acid contact potentials from their frequencies of occurrence in proteins: a lattice model study. *J Phys Condens Matter* 16:2551–2564.

Tobi, D., and R. Elber. 2000. Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins* 41:40–46.

Tuffery, P., C. Etchebest, S. Hazout, and R. Lavery. 1991. A new approach to the rapid determination of protein side chain conformations. *J Biomol Struct Dyn* 8:1267–89.

Vendruscolo, M., R. Najmanovich, and E. Domany. 2000. Can a pairwise contact potential stabilize native protein folds agaionst decoys obtained by threading? *Proteins* 38:134–148.

Vitkup, D., P. Kharchenko, and A. Wagner. 2006. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol* 7:R39.

Wako, H., and T. L. Blundell. 1994a. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. i. solvent accessibility classes. *J Mol Biol* 238:682–92.

Wako, H., and T. L. Blundell. 1994b. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. ii. secondary structures. *J Mol Biol* 238:693–708.

Wald, A. 1949. Note on the consistency of maximumm likelihood. *Ann Math Stat* 20:595–601.

Wall, D.P., A.E. Hirsh, H.B. Fraser, J. Kumm, G. Giaever, M.B. Eisen, and M.W. Feldman. 2005. Functional genomic analysis of the rates of protein evolution. *P Natl Acad Sci USA* 102:5483–5488.

Wang, G., and Jr. Dunbrack, R. L. 2003. Pisces: a protein sequence culling server. *Bioinformatics* 19:1589–91.

Wernisch, L., S. Hery, and S. J. Wodak. 2000. Automatic protein design with all atom force-fields by exact and heuristic optimization. *J Mol Biol* 301:713–736.

Whelan, S., P. Liò, and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet* 17:262 – 272.

Whelan, Simon, and Nick Goldman. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol Biol Evol* 18:691–699.

Wilke, CO, and DA Drummond. 2006. Population genetics of translational robustness. *Genetics* 173:473–481.

Williams, S. G., and S. C. Lovell. 2009. The effect of sequence evolution on protein structural divergence. *Mol Biol Evol* 26:1055–1065.

Wilson, A. C., S. S. Carlson, and T. J. White. 1977. Biochemical evolution. *Annu Rev Biochem* 46:573–639.

Wilson, M. A., and A. T Brunger. 2000. The 1.0 angstrom crystal structure of ca2+-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. *J Mol Biol* 301:1237–1256.

Wolf, M., Y. Wolf, and E. V. Koonin. 2008. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biology Direct* 3:40.

Wright, S. I., C. B. K. Yau, M. Looseley, and B. C. Meyers. 2004. Effects of Gene Expression on Molecular Evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Mol Biol Evol* 21:1719–1726.

Xia, Y., E. S. Huang, M. Levitt, and R. Samudrala. 2000. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol* 300:171–85.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol Biol Evol* 10:1396–401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. *J Mol Evol* 39:306–314.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11:367–372.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15:568–573.

Yang, Z., and R. Nielsen. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* 25:568–579.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–49.

Yang, Z, and D Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol* 12:451–458.

Yang, Z., and W. J. Swanson. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* 19:49–57.

Yu, J., and J. L. Thorne. 2006. Testing for spatial clustering of amino acid replacements within protein tertiary structure. *J Mol Evol* 62:682–92.

Yuan, Z., T. L. Bailey, and R. D. Teasdale. 2005. Prediction of protein b-factor profiles. *Proteins* 58:905–12.

Zaman, M. H., M. Y. Shen, R. S. Berry, K. F. Freed, and T. R. Sosnick. 2003. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides. *J Mol Biol* 331:693–711.

Zhou, Y., H. Brinkmann, N. Rodrigue, N. Lartillot, and H. Philippe. 2010. A Dirichlet Process Covarion Mixture Model and Its Assessments Using Posterior Predictive Discrepancy Tests. *Mol Biol Evol* 27:371–384.

Appendix I

Fast optimization of statistical potentials for structurally constrained evolutionary models

In this article, an alternative optimization procedure to the one described in chapter 2 is presented. The likelihood score is redefined using a *leave-one-out* argument: only one site of the protein is changed at a time, while the other positions are kept with the state in the native sequence. By adapting the statistical framework of chapter 2 to this new definition, computational times are reduced up to 1,000 times, while obtaining essentially the same parameter values.

This is the method used through all chapter 3, which implied numerous comparisons of very high-dimensional models.

BMC Evolutionary Biology

BioMed Central

Methodology article

Fast optimization of statistical potentials for structurally constrained phylogenetic models

Cécile Bonnard*[1,2], Claudia L Kleinman[2], Nicolas Rodrigue[3] and Nicolas Lartillot[2]

Address: [1]Département d'Informatique, LIRMM, 161 rue Ada, 34392 Montpellier Cedex 5, France, [2]Département de Biochimie, Université de Montréal, Montréal, Québec, Canada and [3]Department of Biology, University of Ottawa, Ottawa, Ontario, Canada

* Corresponding author

Published: 9 September 2009

BMC Evolutionary Biology 2009, **9**:227 doi:10.1186/1471-2148-9-227

Received: 9 April 2009
Accepted: 9 September 2009

This article is available from: http://www.biomedcentral.com/1471-2148/9/227

Abstract

Background: Statistical approaches for *protein design* are relevant in the field of molecular evolutionary studies. In recent years, new, so-called structurally constrained (*SC*) models of protein-coding sequence evolution have been proposed, which use statistical potentials to assess sequence-structure compatibility. In a previous work, we defined a statistical framework for optimizing knowledge-based potentials especially suited to SC models. Our method used the maximum likelihood principle and provided what we call the *joint* potentials. However, the method required numerical estimations by the use of computationally heavy *Markov Chain Monte Carlo* sampling algorithms.

Results: Here, we develop an alternative optimization procedure, based on a *leave-one-out* argument coupled to fast gradient descent algorithms. We assess that the leave-one-out potential yields very similar results to the joint approach developed previously, both in terms of the resulting potential parameters, and by Bayes factor evaluation in a phylogenetic context. On the other hand, the leave-one-out approach results in a considerable computational benefit (up to a 1,000 fold decrease in computational time for the optimization procedure).

Conclusion: Due to its computational speed, the optimization method we propose offers an attractive alternative for the design and empirical evaluation of alternative forms of potentials, using large data sets and high-dimensional parameterizations.

Background

Recent advances in computer science and in the acquisition of new genetic sequences from a variety of organisms have opened up a wide spectrum of new possibilities in molecular evolutionary modeling. In particular, codon substitution models explicitly formulated in terms of a balance between mutation and selection constitute an attractive strategy [1-4]. By deriving the substitution process from basic principles of population genetics, their aim is to bridge the gap between population genetics and phylogenetics, and thus to offer a better understanding of the driving forces of the long term evolutionary process. More specifically, these mutation-selection models propose that

the substitution rate from a sequence s to another s' ($R_{ss'}$) depends on the rate of mutation from s to s' ($Q^{mut}_{ss'}$), and on the probability for this mutation to be fixed in the population ($p_{fix}(ss')$):

$$R_{ss'} = Q^{mut}_{ss'} \cdot p_{fix}(ss'). \quad (1)$$

The mutation matrix $Q^{mut}_{ss'}$ depends only on the underlying mutation model, and is generally assumed to be fixed along the lineages and uniform along the sequence. The fixation probability $p_{fix}(ss')$ depends on the particular model chosen.

Among the mutation-selection codon models, we focus on the structurally constrained (SC) models [4-7] which attempt to explicitly link a protein's tertiary structure to the evolution of its sequence. They consider that a protein is under a purifying selection maintaining a stable and constant tertiary structure. Importantly, and unlike most probabilistic models currently used in molecular evolutionary studies, SC models are explicitly site-interdependent, and therefore, require complex Monte Carlo methods to be implemented and applied to empirical data [3,4,8].

In SC models, the fixation probability of a given mutation depends on a score function assessing the adequacy of a sequence s to the tertiary structure of the protein, c. This score should be devised so that the fixation probability is low if the proposed mutation destabilizes the structure or complicates the folding process. Since Anfinsen's experiments [9], the relations between protein structure and sequence have been carefully studied and an intuitive approach consists in relying on first principles of protein thermodynamics, using all-atom force fields (e.g. AMBER [10], CHARMM [11]). However, in our case, the instantaneous rate of substitution ($R_{ss'}$), and thus the structure/sequence score function, have to be computed for each possible nearest neighbor mutant, and for each substitution, along the entire evolutionary tree. Therefore, we need a fast computation of the fixation probability which precludes the use of all-atom force fields.

An attractive alternative is provided by knowledge-based (or statistical) potentials. They mimic the Boltzmann law [12-15] and usually rely on a coarse-grained description of the structure, implicitly integrating out the degrees of freedom of the side chains and thus avoiding the complexity and the computation requirements of all-atom force fields [16-23]. In addition, they are trained empirically from databases of natural proteins. This latter point is of particular interest in evolutionary studies, where we are interested in all aspects of the relations between sequence and structure prevailing in natural sequences, and not only in the specific problem of the thermodynamic stability. In this respect, one expects that learning potentials from native structure-sequence databases using blind machine learning methods will capture all such aspects.

Many statistical potentials have been proposed [12,14,15,19,24,25], either to predict the fold of a given sequence (*protein folding*) or to find a sequence or a set of sequences folding into a given tertiary structure (*protein design*). However, the same potential may not be best-suited to both goals since the spaces of optimization are very different: in the protein folding problem the search is done over the structure space, while in the protein design problem the search is done over the sequence space. The phylogenetic context described here is more akin to a protein design perspective, as the structure of the protein is assumed constant during evolution, representing a constraint under which the sequence is evolving.

Several methods have been developed to train statistical potentials in a protein design perspective [19,24,25]. In a previous work, we introduced a probabilistic framework for protein design purposes based on the maximum likelihood principle [26]. The likelihood we considered was the probability of the sequences S given their native structures C and the model parameters (here, the statistical potential parameters, θ), $P(S|C, \theta)$. This probability was then maximized with respect to the potential parameters (e.g. pairwise contact energy coefficients) by a gradient method. However, the probability $P(S|C, \theta)$ involves a normalizing factor, summing over all possible sequences, which cannot be analytically calculated. We thus had to resort to a Markov Chain Monte Carlo (*MCMC*) numerical procedure: at each step of the gradient descent, we generated a set of sequences by Gibbs sampling, conditional on the current values of the potential. This set of sequences was then used to estimate the gradient. The Gibbs sampling procedure was the limiting step of our algorithm, restricting the set of alternative potentials that we could explore and empirically test. The potentials we obtained using this method are called *joint* potentials hereafter.

Interestingly, Kuhlman and Baker [27] used a *leave-one-out* procedure to estimate a restricted set of parameters of a free physical energy function in order to do protein design. In this procedure, only one site of the protein is changed at a time, while the other positions are kept fixed in their native state. The procedure is thus similar to training a potential to recognize acceptable sequence variants, given the target structure, among all possible point mutants. The leave-one-out criterion seems to give good results. However, it has never been assessed against alternative methods. Here, we adapt the statistical framework

we defined in [26] now using the leave-one-out definition of the likelihood to perform the gradient descent instead of the joint likelihood. We compare the potential parameters obtained by the two methods, and we establish that we can be highly confident in the results obtained using the leave-one-out likelihood. Overall, the leave-one-out procedure allows much faster computations while giving sensibly the same results as the joint one.

Results

Likelihood framework

As in [26], we formulate the problem in terms of a probabilistic model, considering a sequence $s = (s_i)_{1..n}$ of length n according to a probability distribution $P(s|c, \theta)$, conditional on the conformation c and on a set of potential parameters θ. The parameters are estimated by maximizing the probability of observing a database of N independent sequence-structure pairs $(\tilde{S}, C)$, with $\tilde{S} = (\tilde{s}^p)_{p=1..N}$, $C = (c^p)_{p=1..N}$. Here, $\tilde{s}^p = (\tilde{s}_i)_{i=1..n_p}^p$ is the p-th native sequence of the dataset, n_p is the lenght of this sequence and c^p is the native conformation associated with $\tilde{s}^p$. In practice, a native sequence-structure pair corresponds to a protein taken from the PDB.

The probability that we want to maximize can be expressed as follows:

$$P(\tilde{S} \mid C, \theta) = \prod_p P(\tilde{s}^p \mid c^p, \theta). \quad (2)$$

As a function of θ, this term can be seen as a likelihood. Hereafter, we define the methodology with one protein, but it can be easily generalized to a set of proteins.

Borrowing from [26], we set:

$$P(s \mid c, \theta) = \frac{e^{-G(s|c,\theta)}}{\sum_{s' \in \mathbb{S}} e^{-G(s'|c,\theta)}} = \frac{e^{-G(s|c,\theta)}}{Y}, \quad (3)$$

where Y is called the *normalization factor*, and $G(s|c, \theta)$ the *inverse potential*, defined as

$$G(s \mid c, \theta) = E(s \mid c, \theta) - F(s), \quad (4)$$

where $E(s|c, \theta)$ is the statistical potential and $F(s)$ is analogous to a free energy term and can be approximated using the *random energy model* [19,28-30]:

$$F(s) = \sum_{1 \le i \le n} \mu_{s_i}, \quad (5)$$

where μ_a, $a = \{1..20\}$ are unknown parameters, analogous to *chemical potentials* [26].

Optimization method

Joint likelihood maximization

In our previous work [26], we defined a score function $\omega(\tilde{s} \mid c, \theta)$ as:

$$\omega(\tilde{s} \mid c, \theta) = -\ln P(\tilde{s} \mid c, \theta) = G(\tilde{s} \mid c, \theta) + \ln Y. \quad (6)$$

This score function should be minimized conditional to θ. Its gradient is:

$$\frac{\partial \omega(\tilde{s}|c,\theta)}{\partial \theta} = \frac{\partial G(\tilde{s}|c,\theta)}{\partial \theta} + \frac{\partial \ln Y}{\partial \theta} = \frac{\partial G(\tilde{s}|c,\theta)}{\partial \theta} - \left\langle \frac{\partial G}{\partial \theta} \right\rangle, \quad (7)$$

where $\langle \cdot \rangle$ stands for the expectation over sequences drawn from the probability defined by eq. 3. Given the size of the sequence space (20^n), this expectation cannot be computed analytically, and therefore, in [26] we used a MCMC method to numerically estimate this expectation.

Leave-one-out likelihood maximization

We define for site i, $i = 1..n$, the leave-one-out probability

$$P_i^l(s_i = a \mid \tilde{s}_{\setminus i}, c, \theta) = P_i^l(s_i = a \mid \forall j \neq i \; s_j = \tilde{s}_j, c, \theta), \quad (8)$$

which is the probability of having an amino acid a at site i, in the context of the native sequence at all other sites ($\forall j \neq i \; s_j = \tilde{s}_j$). This leave-one-out probability can easily be obtained by a normalization over all possible twenty outcomes at site i:

$$P_i^l(s_i = a \mid \tilde{s}_{\setminus i}, c, \theta) = \frac{e^{-G_i(s_i=a|\tilde{s}_{\setminus i},c,\theta)}}{\sum_{k=1}^{20} e^{-G_i(s_i=k|\tilde{s}_{\setminus i},c,\theta)}}. \quad (9)$$

We can write this probability for any amino acid a, and in particular for the native amino acid at site i, $\tilde{s}_i$ i.e. $P_i^l(s_i = \tilde{s}_i \mid \tilde{s}_{\setminus i}, c, \theta)$. Taking the product over all positions $i = 1..n$, and by analogy with our previous definition of likelihood, we define the leave-one-out likelihood:

$$P^l(\tilde{s} \mid \tilde{s}, c, \theta) = \prod_{1 \le i \le n} P_i^l(s_i = \tilde{s}_i \mid \tilde{s}_{\setminus i}, c, \theta). \quad (10)$$

Note that this leave-one-out likelihood is normalized over the sequences, exactly as in the case of eq. 3. Therefore it yields a valid probability distribution over the sequence

space. On the other hand, the probability depends not only on c and θ, but also, in some sense, on the native sequence itself. To make this point explicit, we make $\tilde{s}$ appear on both sides of the conditioning bar.

We define the corresponding scoring function:

$$\omega^l(\tilde{s} \mid \tilde{s}, c, \theta) = -\ln P^l(\tilde{s} \mid \tilde{s}, c, \theta), \tag{11}$$

the gradient of which is immediately obtained (Additional File 1):

$$\frac{\partial \omega^l(\tilde{s}|\tilde{s},c,\theta)}{\partial \theta} = \sum_{i=1..n} \frac{\partial G_i(s_i=\tilde{s}_i|\tilde{s}\backslash_i,c,\theta)}{\partial \theta} - \sum_{i=1..n} \sum_{a=1..20} p_i(a) \frac{\partial G_i(s_i=a|\tilde{s}\backslash_i,c,\theta)}{\partial \theta}. \tag{12}$$

This gradient can be analytically calculated, at each step of a gradient descent. We note that the term corresponding to the normalization factor (the second term in eq. 12) can be seen as an expectation over the leave-one-out probability. It is thus analogous to the expectation appearing in the right hand of eq. 7. However, it is defined on a much more restricted universe ($20 \cdot n$ states, compared to the 20^n states in the case of the joint likelihood).

For implementing both methods, we used a simple form of potential [26], consisting in two terms: one related to contact interactions and the other to the solvent accessibility (see Methods).

Potential optimization

We first run our leave-one-out method on DS_l (see Methods). We consider that the optimization is complete when the overall maximum gradient is smaller than 10^{-2}. This corresponds to a variation of 10^{-6}, at most, in the value of the potential parameters. Using this stopping condition on the dataset DS_l with empirically tuned general steps (e.g for the contact parameters: $\delta^c_{grad} = 10^{-5}$ and for the solvent accessibility parameters: $\delta^a_{grad} = 10^{-4}$), we compare three different gradient descent methods (described in Methods): the simple gradient descent, the inertial gradient descent, and the controlled inertial gradient descent. The values of the parameters stabilized after 14,500 gradient steps for the simplest gradient descent, versus 1,500 gradient steps for the inertial gradient, and 1,200 gradient steps for the controlled inertial gradient. Concerning the last method, if we choose a different general step (e.g. $\delta^a_{grad} = 10^{-3}$ and $\delta^c_{grad} = 10^{-2}$) the procedure automatically reaches the optimal step for that dataset. At the beginning of the optimization procedure, the inertial component of the gradient greatly speeds up the optimization, but is automatically deactivated when the values of the potential parameters are near the optimum, thus avoiding the numerical instabilities usually observed using less adaptive gradient methods.

Independent runs from different and randomly chosen initial values for the parameters of the leave-one-out potential (θ), lead to the same final values of $\omega^l(\tilde{s} \mid \tilde{s}, c, \theta)$ (fig. 1) and of the potential parameters (fig. 2). These computations were done with the three gradient descent methods, and resulting always in the same final values, which suggests that, in the present case, we do not have local minima in the space of parameters. Similarly, the potential parameters obtained by two independent runs on the same dataset are very similar, indicating that our stopping condition is sufficient to have a good precision in our estimates (Additional file 2). In fig. 1 we have also represented the evolution of some parameters of the potential during optimization. We can see that the values of these parameters oscillate at the beginning of the gradient descent and then reach their optimal values. This behavior is caused by the evolution of the other parameters, as they influence each other during optimization. The complete series of parameter values obtained by our optimization method are presented in the additional file 3.

The contact potentials obtained with the leave-one-out optimization criterion make sense from a biological point of view (fig. 3): as expected, favorable interactions between amino acids in the contact potentials are represented by large negative value (e.g. the Cysteine-Cysteine contact energy, fig. 3), and by large positive value for unfavorable interactions (e.g. the Lysine-Lysine or Lysine-Arginine interactions, which are electrostatically repulsive). Concerning the accessibility potentials, it is important to note that we are working in a protein design context (i.e. we are evaluating the fitness of alternatives amino acids in a given accessibility class). Accordingly, the accessibility potentials have to be interpreted rowwise. If one wants to compare the accessibility potentials between classes for a given amino acid (i.e. in a protein folding perspective), one solution is to remove the logarithm of the frequency of the accessibility classes to each potential (additional file 4). Also, note that there is a lack of identifiability between α and μ, which has been be resolved by including the chemical potentials in the accessibility terms.

Complexity

In our previous work, we had to use a MCMC protocol to numerically evaluate the derivative of the gradient (see.

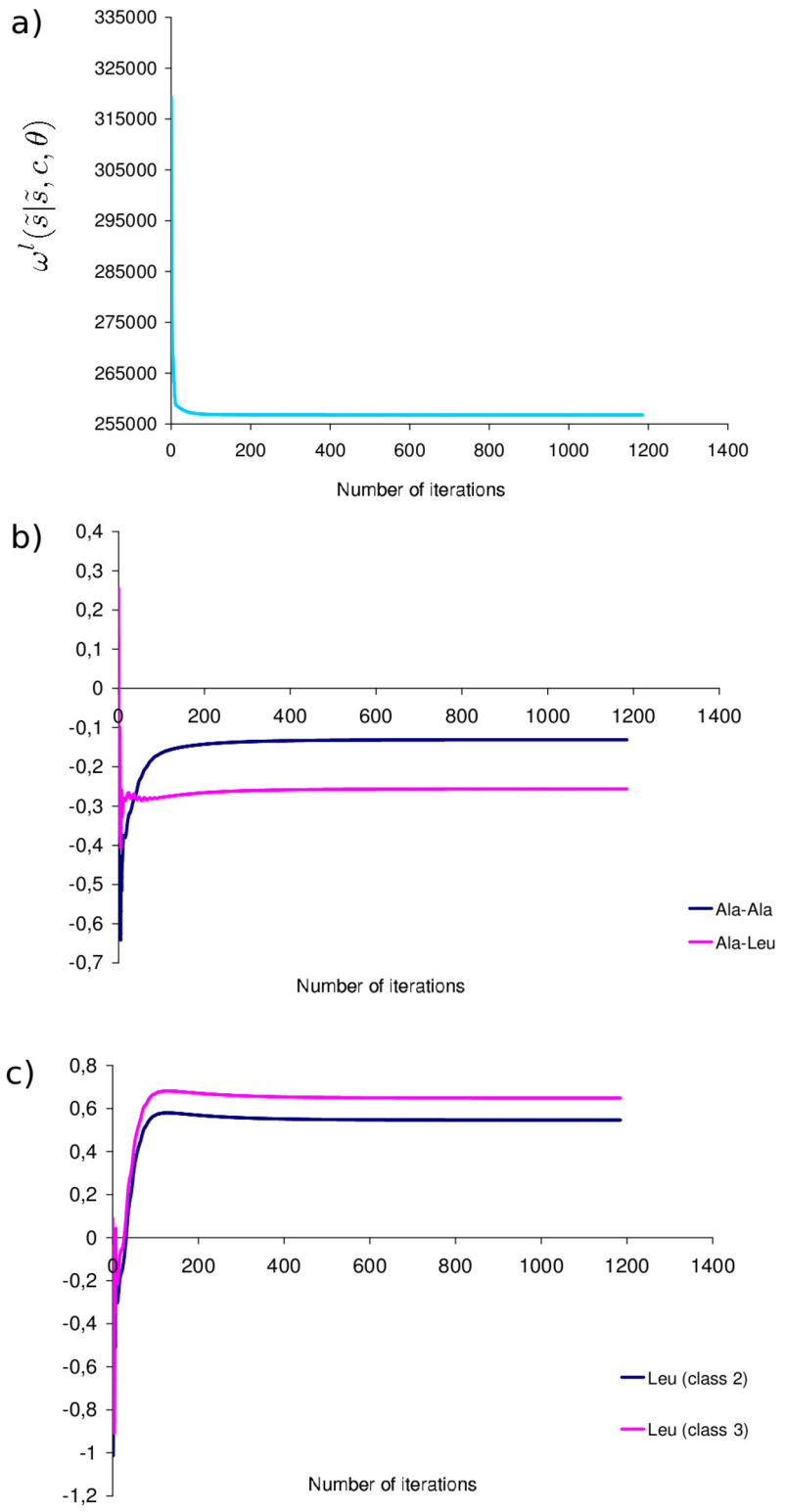


Figure 1

Convergence of the optimization procedure. Evolution of (a) the score function, (b) contact potential parameters and (c) accessibility potential parameters, for the dataset DS_l, using the controlled inertial gradient descent.

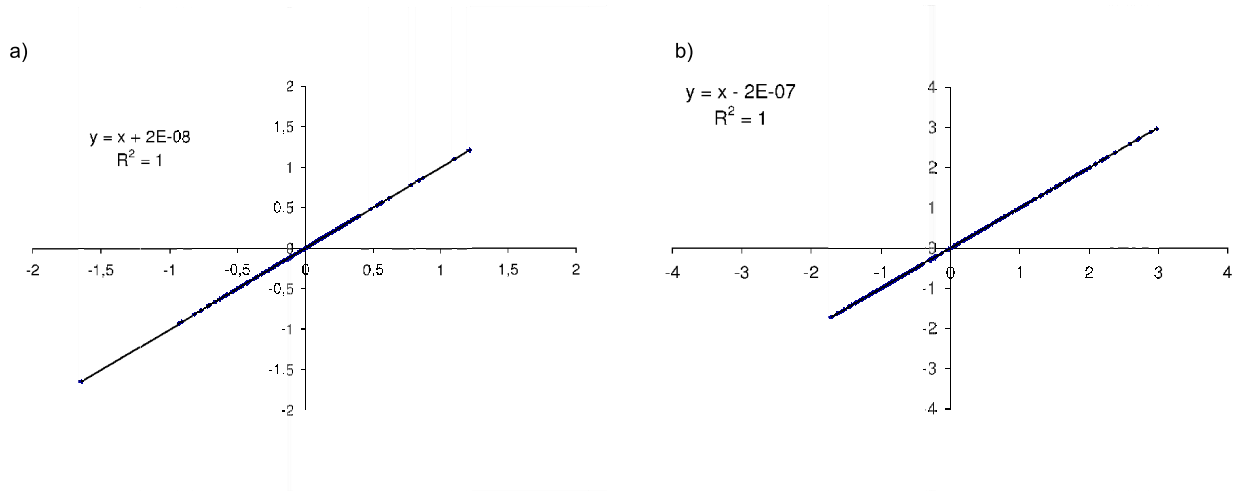


Figure 2
XY comparisons of the leave-one-out potential parameters. XY comparisons of two independent runs on the same dataset DS_l for (a) contact and (b) solvent accessibility potential parameters respectivly.

eq. 7), which was a computationally demanding task. At each step of the gradient descent, we had to sample a set of sequences by Gibbs sampling, under the current values of the parameters, so as to numerically estimate the gradient of the log-likelihood.

To compare the joint and the leave-one-out potentials, we first define an elementary calculation as the evaluation of the *inverse* potential at a particular site i for one particular amino acid a (what we called $G_i(s_i = a|\,\tilde{s}_{\setminus i}, c, \theta)$, eq. 9). This calculation has to be made in both cases. It is explicitly defined in the leave-one-out procedure (eq. 10), and is implicitly used in the joint context: an elementary step of the Gibbs sampling algorithm consist in computing, at a given site i the leave-one-out probability (eq. 9) for each possible amino-acid at this site, conditional on the rest of the sequence, and to choose the new aminoacid at site i according to these probabilities. Performing such an elementary update for every site in turn corresponds to one Gibbs sampling sweep and represents $20 \cdot n$ elementary computations. A reliable estimate of the joint expectation requires K sweeps (burn in included) and so, for a gradient step, we need $K \cdot n \cdot 20$ elementary calculations (in practice, $K \simeq 1{,}000$).

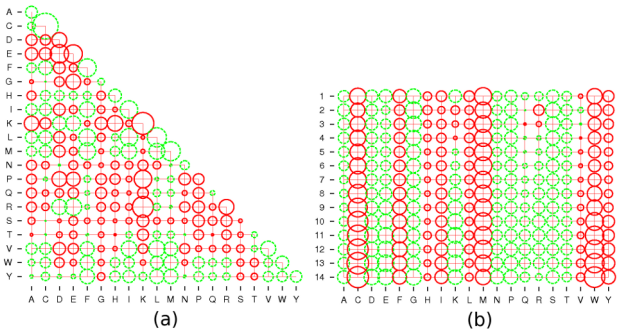


Figure 3
Validation of the potential parameters. Bubble plot representations of (a) contact potential parameters and (b) accessibility potential parameters obtained upon the dataset DS_l. Negative values are plotted in green while positive values are plotted in red.

In the case of the leave-one-out potential, we only have to make the equivalent of one sweep to exactly compute the gradient (eq. 12). Thus, we only need $n \cdot 20$ elementary calculations for a gradient step, which thus represents a 1,000-fold increase in computational speed compared to the joint method. In practice, and after the addition of the acceleration of the gradient descent, it took about one week to have a good estimate when we used the joint method, versus less than fifteen minutes when using the leave-one-out approach.

Potentials are indistinguishable

We applied the two optimization procedures (joint and leave-one-out) to the same dataset DS_j, and found a high correlation between the two resulting potentials (fig. 4). The correlation coefficient R^2 was about 0.96779 for the contact potential parameters and about 0.97374 for the accessibility potential parameters. For comparison, we applied the leave-one-out procedure on the two datasets DS1 and DS2 (see additional file 2) and found a correlation coefficient of 0.9477 for the contact parameters and

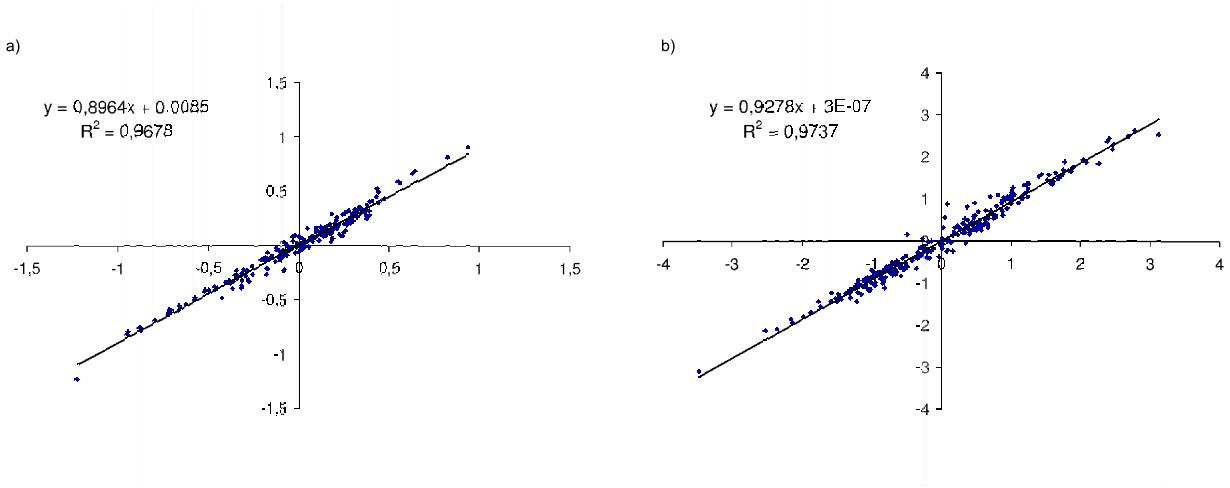


Figure 4
XY comparisons of the leave-one-out and joint potential parameters. XY comparisons between the two potentials (optimized on the same dataset DS_j), with, in X-axis the leave-one-out potential, and in Y-axis the joint potential. (a) represents the correlation between the contact potential parameters, and (b) the correlation between the accessibility potential parameters.

of 0.9596 for the accessibility parameters, indicating that the difference between the joint and the leave-one-out potentials is small compared to the sampling error due to the finite size of the training set. Altogether, the leave-one-out method appears to be a fast and reliable optimization procedure, yielding potentials that are virtually indistinguishable from those obtained under the joint method. As presented in [26], the contact potentials present a correlation (R^2 = 0.6565) with those of Miyazawa and Jernigan [13].

Phylogenetic evaluation

In eq. 1, we defined the substitution process of the SC model as a process depending on a mutation rate and a fixation probability. There are many ways the fixation probability could be expressed. Here, we do as in Robinson et al [4] and assume that this probability depends only on the potential difference (ΔG) between the original and the mutated sequences. Let us denote by s_{nuc} and s'_{nuc}, two sequences which differ only by a nucleotide, and s_{aa} and s'_{aa}, the corresponding amino acid sequences (which may be identical due to codon synonymy). Then, the rate of substitution between s and s' is:

$$R_{s_{nuc}s'_{nuc}} = Q^{mut}_{s_{nuc}s'_{nuc}} \cdot e^{-\beta\Delta G_{s_{aa}s'_{aa}}}, \qquad (13)$$

where $Q^{mut}_{s_{nuc}s'_{nuc}}$ is the mutation term depending only on the two sequences s_{nuc} and s'_{nuc}. $\Delta G_{s_{aa}s'_{aa}}$ is the energy difference between s_{aa} and s'_{aa}, and $\beta \geq 0$ can be considered as the strength of the structure-sequence constraint enforced by the model. Thus, a negative (resp. positive) ΔG means that the mutation is more (resp. less) likely to be accepted than a purely neutral (e.g. synonymous) mutation.

Note that the substitution process defined by eq. 13 is reversible and has a stationary distribution defined by:

$$\Pi_s \propto \Pi_0(s_{nuc})e^{-2\beta G(s_{aa})}, \qquad (14)$$

where $\prod_0(s_{nuc})$ is the stationary distribution induced by the pure mutation process ($Q^{mut}_{s_{nuc}s'_{nuc}}$). Given the way our potentials are optimized (see eq. 3 and 9) and assuming that natural sequences are sampled at equilibrium from the process defined by eq. 13, we then expect that the optimal value of β should be close to 0.5. In the following, we explore the entire range $\beta \in [0, 1]$.

We denote by SC^l_β the SC model defined using the leave-one-out potential and SC^j_β the SC model defined using the joint potential; the two models depend on β. Obviously, when $\beta = 0$, $SC^l_0 = SC^j_0 = SC_0$, and the model reduces to a pure mutation model which will be considered as our reference model.

We implemented our potential in the SC model as described in [3] and applied it to the GLOBIN15-144 dataset, with an underlying mutational specification inspired by the codon model in [31] and denoted as MG in [3]. This MCMC framework allows one to obtain a sample of parameter values and substitutional histories along the tree, drawn from the posterior distribution under the $SC^{l}_{0.5}$ model. Such a sample can then be marginalized over quantities of interest. Here, we briefly illustrate the approach by displaying the logo of the reconstructed mammalian ancestor hemoglobin sequence (fig. 5).

Since the leave-one-out procedure can be seen as an approximate but faster training method, compared to the joint method developed previously, we evaluated its impact on model fit via Bayes factors evaluations (see Methods). In this section we consider the three versions of the SC model, SC^{l}_{β}, based on a contact + accessibility leave-one-out potential, SC^{j}_{β}, based on a contact + accessibility joint potential, and SC^{c}_{β} based on a contact only joint potential. As explained in the methods, in the present case, the thermodynamic integration method yields a complete fitness curve (fig. 6) of each model (i.e. a curve representing the Bayes factor of each model against the reference model, as a function of β). In this way, we can readily spot the optimal value of β under each model, and report the Bayes factors under this optimal value (table 1).

As can be seen from fig. 6 and table 1, the models based on the joint and the leave-one-out potentials have a very similar fit across the whole range of value of β that we tested. Interestingly, in all but one cases, the Bayes factor

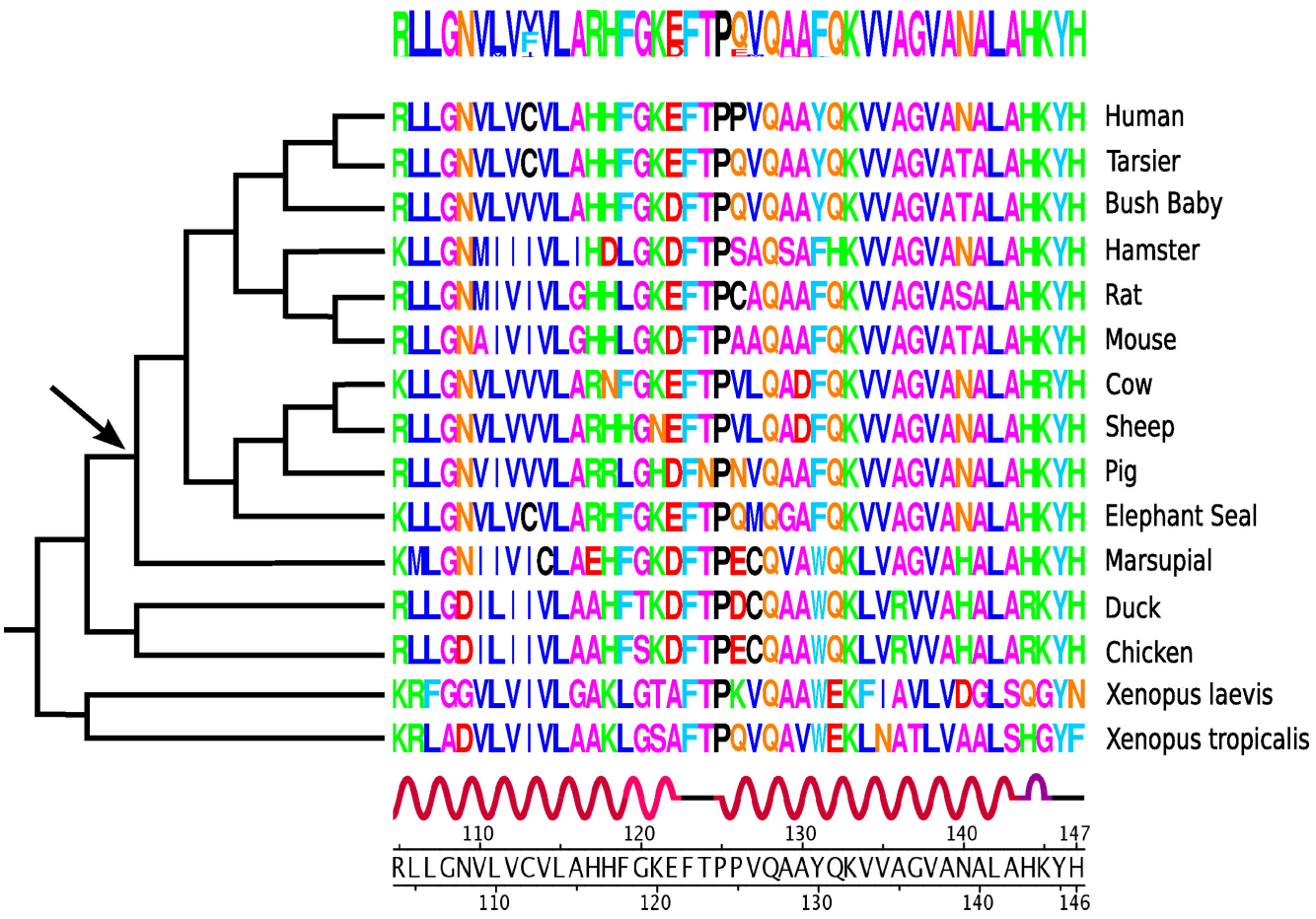


Figure 5
Logo profile of the mammalian ancestral globin sequence. The node is marked by an arrow. The translated sequences of the true alignment are displayed along with the secondary structure of the structure PDB code 4HHBB.

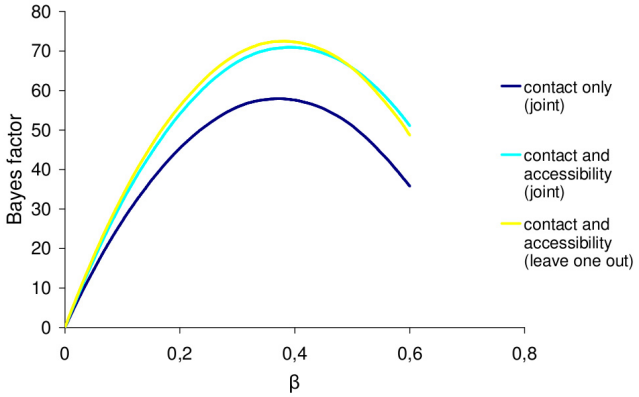


Figure 6

Bayes factor. Curves representing the Bayes factor as a function of β, with SC^{l}_{β} (in yellow), SC^{j}_{β} (in light blue) and SC^{c}_{β} (in dark blue), for the dataset BGLOBIN15-144.

appears to be slightly in favor of the leave-one-out potential, although the differences are not significant. As a point of comparison, we also measured the fit of the contact only potential (joint method), to illustrate that the difference between the joint and the leave-one-out methods is small compared to the differences observed between the alternative forms of statistical potential that we would like to empirically compare (see [26] for an evaluation of the relative contribution of each potential component to the fitness of the model).

Discussion

In a previous work [26], we defined a statistical framework for protein design, using the maximum likelihood principle, with the aim of devising statistical potentials to be used in phylogenetic studies. However, the optimization procedure we introduced at that time requires a MCMC protocol to cope with the proportionality constant entailed by the normalization of the probability over the sequence space. Here, we introduce a different likelihood, which we called leave-one-out, to optimize the potentials. A similar procedure was previously used by Kuhlman and Baker [27], but was not statistically assessed against alternative procedures. We found in this work that the joint and the leave-one-out potentials are virtually indistinguishable, both by direct comparison and by Bayes factor evaluation in a phylogenetic context.

We note that the optimal β for the SC^{l}_{β} model is not 0.5, as one may expect given the way our potentials were normalized (see eq. 3, 6 and 13). Several explanations can be proposed. First, strictly speaking, this expectation is valid under the joint procedure, and not under the leave-one-out procedure. But the very high similarity between the two resulting potentials, and the fact that a similar phenomenon ($\beta \neq 0.5$) can be observed also under a potential optimized using the joint method [3] do not favor this explanation. Alternatively, it may appear at first that this could be due to the fact that the underlying mutation model (the Q^{mut} matrix in eq. 13) was not explicitly taken into account when optimizing the potential (so that the chemical potentials implicitely include a mutational component), whereas our phylogenetic model does involve an explicit mutational process. In this sense, in the phylogenetic analysis, there is a potentially (partially) redundant modeling of mutational features, in having explicit parameters devoted to these, in combination with the use of the SC setting. This might explain the optimal value of β lower than 0.5. The phenomenon may also be the result of model violations, which are very likely to be present given the simple form of the potentials. Finally, it is also likely that the mutation pressure, or the selection strength (represented by β) is not the same for each protein. Accordingly, two possible improvements to the method can thus be proposed here: the first is to optimize the potential while allowing for different values of β for each protein or each family of protein. The second is to cluster proteins into classes, and optimize a potential specifically for each class.

Table 1: The natural logarithm of the Bayes factors.

	ADH23-254	CALM36-444	GLOBIN15-144	LYS25-134
SC^{ι}_{β}	[74.748-75.032]	[149.819-149.929]	[57.953-58.135]	[11.5-11.968]
SC^{J}_{β}	[102.666-102.766]	[161.340-161.491]	[70.666-70.948]	[26.287-26.417]
SC^{ι}_{β}	[102.977-103.115]	[158.679-158.858]	[72.485-72.872]	[29.545-29.852]
optimal β	[0.387-0.397]	[0.371-0.383]	[0.450-0.498]	[0.179-0.249]

Conclusion

Apart from these two possible improvements, many other directions of research should now be explored: alternative functional forms for the potential should be implemented and empirically tested. Several methods accounting for negative design, through the use of explicit decoys [18] such as the use of a normalized energy gap between a native structure and misfolded structures [32], or using variational methods [19], also deserve further investigation. The supervised learning described here depends on structure-sequence pairs. In the present case, we have used native pairs, but this could be relaxed by taking a set of structures (e.g. obtained by molecular dynamics) as the reference structure or by taking a set of homologous sequences instead of a unique sequence [33]. A more appealing method would consist in doing the optimization directly within the phylogenetic context. Importantly, the fact that the leave-one-out procedure is much faster than the joint method (in the present case, roughly by a factor 1,000), has obvious practical consequences, as it allows a much larger diversity of alternative models and methods to be tested.

Methods

Gradient descent

When performing a gradient descent, several methods can be used. We expose here the three gradient descent methods that we compared. In all cases, the method rely on a cyclical updating of parameter values, where, given the values of parameters at the m^{th} cycle, which we write as $\theta^{(m)}$, the update is given by:

$$\theta^{(m+1)} = \theta^{(m)} - \Delta\theta^{(m+1)}. \quad (15)$$

The increment, $\Delta\theta^{(m+1)}$, is conditional to the scoring function, that we simply denote in this part as $\omega(\theta^{(m)})$.

Fixed step gradient

This is the simplest form of the gradient descent. We write:

$$\Delta\theta^{(m+1)} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m)})}{\partial\theta}, \quad (16)$$

where δ_{grad} is the fixed step of the gradient descent. Even though this formalism is simple, the choice of the step is not trivial. Indeed, if the step is too large, the values of the potential will oscillate around the optimal values. Conversely, if the step is too small, the gradient descent will be too slow.

Inertial gradient

To reduce the optimization time, another method of gradient descent was developed, based on an analogy with the physical phenomenon of inertia.

$$\Delta\theta^{(m+1)} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m)})}{\partial\theta} + \delta_{iner} \cdot \Delta\theta^{(m)}. \quad (17)$$

δ_{iner} is the damping rate of the inertial component, $0 \leq \delta_{iner} < 1$. If $\delta_{iner} = 0$, eq. 17 reduces to the case of the simple gradient. In practice, we set δ_{iner} equal to 0.9.

However, there is a drawback when taking into account the previous variation of the parameters: when the directions of the gradient change, the inertial part of the gradient brings the parameters too far beyond the maximum. In addition, the gradient step δ_{grad} has to be small enough so that the values of the potential do not oscillate around the optimal values, as in the case of the fixed step gradient.

Controlled inertial gradient

To avoid these two drawbacks, we define here a controlled inertial gradient descent formalism. Specifically, let us define:

$$\Delta\theta^{*} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m-1)})}{\partial\theta} + \delta_{iner} \cdot \Delta\theta^{(m)}, \quad (18)$$

$$\Delta\theta^{\bullet} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m-1)})}{\partial\theta}. \quad (19)$$

The decision procedure can thus be described as follows (see additional file 5). First, we test if the addition of $\Delta\theta^{*}$ (derivative component and inertial component) to the actual values of parameters $\theta^{(m)}$ gives a higher likelihood than $\theta^{(m)}$. If it does, then the step corresponds to a classical step of the inertial gradient descent. Otherwise, the algorithm tests if the addition to $\theta^{(m)}$ of the derivative component ($\Delta\theta^{\bullet}$) only gives a higher likelihood than the actual values. If it does, the step corresponds to a classical gradient descent. Otherwise, we retry a simple gradient descent with a smaller δ_{grad}.

The above procedure has two advantages. The first is the speed-up offered by the inertial component, when its addition has a positive influence on the likelihood. The second advantage is that the last part of the algorithm automates the search for an optimal value of the steps, and avoids both oscillations of θ around the optimum, and a slow gradient descent.

Statistical potentials

We used the same statistical potential function as in our previous work [26]. The (pseudo) energy score consists of two terms:

$$E(s \mid c) = \sum_{1 \leq i \leq j \leq n} \Delta_{ij} \varepsilon_{s_i s_j} + \sum_{1 \leq i \leq n} \alpha_{s_i}^{v_i}. \quad (20)$$

The first term represents the contact free energy (defined between sidechain centers): $\Delta_{ij} = 1$ if i and j are closer than the cutoff distance (here 6.5 Å), and ε_{ab} represents the contact potential between amino acids a and b. The second term represents the accessibility free energy: ν_i is the accessibility class of the site i and α_a^d is the solvent accessibility potential of the amino acid a when placed into the accessibility class d ($d = \{1..D\}$), where D is the number of accessibility classes.

We use the *random energy model* principle to approximate $F(s)$ (eq. 5), so that the inverse potential becomes:

$$G(s \mid c, \theta) = \sum_{1 \le i \le j \le n} \Delta_{ij} \varepsilon_{s_i s_j} + \sum_{1 \le i \le n} \alpha_{s_i}^{\nu_i} + \sum_{1 \le i \le n} \mu_{s_i}. \tag{21}$$

As in our previous work we fix the constraints:

$$\sum_{1 \le a \le 20} \mu_a = 0, \tag{22}$$

$$\sum_{1 \le a \le 20} \sum_{1 \le b \le 20} \varepsilon_{ab} = 0, \tag{23}$$

$$\sum_{1 \le a \le 20} \alpha_a^d = 0, d = \{1..D\}, \tag{24}$$

since $G(s|c, \theta)$ is invariant under the following transformations $\mu_a' = \mu_a + J_1$, $\varepsilon_{ab}' = \varepsilon_{ab} + J_2$ and $\alpha_a'^d = \alpha_a^d + J_3$. However, there is an additional lack of identifiability between a and μ, which can be resolved by including the chemical potentials in the accessibility terms. Indeed, the μ_a terms can be seen as an additive constant to each accessibility term for a given accessibility class (see additional file 6). In the present case, our final inverse potential is therefore:

$$G(s \mid c) = \sum_{1 \le i \le j \le n} \Delta_{ij} \varepsilon_{s_i s_j} + \sum_{1 \le i \le n} \alpha_{s_i}^{\nu_i}, \tag{25}$$

and our set of parameters for the statistical potential will thus consist of:

$$\theta = \{\varepsilon_{ab}, \alpha_a^d\}, \quad 1 \le a \le 20, \quad 1 \le b \le 20, \quad d = \{1..D\}. \tag{26}$$

Bayes factor evaluation

In a Bayesian statistical framework the method of choice for comparing models is to compute Bayes factors. The Bayes factor between two models is defined as the ratio of their respective marginal likelihood. The case $B(SC_0, SC_\beta^l) > 1$ (resp. $B(SC_0, SC_\beta^l) < 1$) is considered as an evidence in favor of (resp. against) the SC_β^l model. We write the Bayes factor between SC_0 and SC_β^l as:

$$B(SC_0, SC_\beta^1) = \frac{P(A|SC_\beta^1)}{P(A|SC_0)}, \tag{27}$$

where A corresponds to the data, composed by an alignment of coding nucleotide sequences and a topology and

$$P(A \mid SC_\beta^l) = \int_\theta P(A \mid \theta) P(\theta) d\theta. \tag{28}$$

Here we compute Bayes factors by thermodynamic integration (or *path sampling*) as described in [3]. The procedure consists in sampling along a continuous path between SC_0 and SC_β^l through a set of slight changes in the value of β. In fact, this procedure provides a complete curve representing the fit of the model as a function of β. Sampling from $\beta = 0$ to $\beta = \beta_{max}$ and from $\beta = \beta_{max}$ to $\beta = 0$ gives two different curves for the logarithm Bayes factor, which we used as an internal check of the reliability of the method (not shown).

Datasets

Optimization datasets

The datasets are made of proteins (structure-sequence pairs) culled from the PDB, with less than 25% of mutual sequence identity and a resolution better than 2 Å [34]. This sequence homology percentage and the size of the database avoid possible bias that could be induced by related proteins. To compare the joint and leave-one-out potentials, we used the dataset on which we previously estimated the joint potentials, DS_j. This dataset is made of 441 proteins and 98,155 sites [26]. We also consider a dataset DS_l (made of 3,363 proteins and 835,717 sites) which was split into two subsets: $DS1$ (1,691 proteins and 419,208 sites), and $DS2$ (1,672 proteins and 416,509 sites). To determine the accessibility classes, we first compute the solvent accessibility area using Naccess 2.1 [35] and partitioned the resulting values into classes [26].

Phylogenetic Datasets

The SC model was applied to 4 distinct multiple sequence alignments: GLOBIN15-144, LYSIN25-134, ADH23-254 and CALM33-444. GLOBIN15-144 is made of 15 vertebrates sequences of the β-globin gene (taken from the original dataset from [36]), with a protein structure defined by the

PDB file 4HHB and a tree topology estimated using Phylobayes 3.1c [37] (which is consistent with the tree topology described in [38]). LYSIN25-134 is made of 25 Abalone sperm lysin sequences [39], with a protein structure defined by the PDB file 1LYS and the tree topology previously defined by [39]. ADH23-254 is made of 23 alcohol dehydrogenase sequences taken form Drosophila [36], with a protein structure defined by the PDB file 1A4U and the tree topology previously defined by [36]. CALM36-444 is made of 36 calmodulin sequences taken from eukaryotes, with a protein structure defined by the PDB file 1CFD and the tree topology estimated using phyML [40] under the model JTT + F + Γ [41,42].

Authors' contributions

CB implemented the leave-one-out and gradient descent methods described here and performed the run of all the experiments. CLK implemented the data pre-processing methods. NR implemented the phylogenetic framework. NL set up the theoretical framework and directed the overall project. All the authors co-wrote the manuscript and approved the final manuscript.

Additional material

Additional file 1

Derivatives of the potential parameters.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S1.pdf]

Additional file 2

XY-comparison of the leave-one-out potentials estimated from two independent datasets: (a) and (b) two independent runs on DS*1 (X-axis) and* DS*2 (Y-axis) for contact and accessibility potentials respectivly.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S2.eps]

Additional file 3

Contact potentials and solvent accessibility potentials written in an alphabetical order.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S3.txt]

Additional file 4

Bubble plot of the solvant accessibility potential where we remove from each potential the corresponding natural logarithm frequency of the accessibility class.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S4.eps]

Additional file 5

Controlled inertial gradient algorithm.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S5.pdf]

Additional file 6

Inclusion of μ_a *in the accessibility terms.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S6.pdf]

Acknowledgements

The authors are grateful to the three anonymous referees for their useful comments on the manuscript. CB was financially supported by the french Centre National de la Recherche Scientifique (CNRS), the Région Languedoc-Roussillon and the Université de Montréal, CLK by NSERC, CIHR and the Université de Montréal, NR by NSERC, and NL by the Université de Montréal, NSERC and the CNRS.

References

1. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15(7):**910-917.
2. Yang Z, Nielsen R: **Mutation-Selection models of codon substitution and their use to estimate selective strengths on codon usage.** *Mol Biol Evol* 2008, **25(3):**568-579.
3. Rodrigue N, Kleinman CL, Philippe H, Lartillot N: **Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons.** *Molecular Biology and Evolution* 2009, **26(7):**1663-1676.
4. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL: **Protein evolution with dependence among codons due to tertiary structure.** *Molecular Biology and Evolution* 2003, **20(10):**1692-1704.
5. Rodrigue N, Lartillot N, Bryant D, Philippe H: **Site interdependence attributed to tertiary structure in protein evolution.** *Gene* 2005, **347(2):**207-217.
6. Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL: **Quantifying the impact of protein tertiary Structure on molecular evolution.** *Mol Biol Evol* 2007, **24(8):**1769-1782.
7. Parisi G, Echave J: **Structural constraints and emergence of sequence patterns in protein evolution.** *Mol Biol Evol* 2001, **18(5):**750-756.
8. Choi SC, Redelings BD, Thorne JL: **Basing population genetic inferences and models of molecular evolution upon desired stationary distribution of DNA or protein sequences.** *Phil trans R Soc B* 2008, **363:**3931-3939.
9. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181:**223-230.
10. Case D, A Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, A KP: *AMBER 10* University of California, San Francisco; 2008.
11. MacKerel AD Jr, Brooks CL III, Nilsson L, Roux B, Won Y, Karplus M: **CHARMM: The Energy Function and Its Parameterization with an Overview of the Program.** In *The Encyclopedia of Computational Chemistry Volume 1*. Edited by: v Schleyer RP, et al. John Wiley & Sons: Chichester; 1998:271-277.
12. Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules* 1985, **18:**534-552.
13. Miyazawa S, Jernigan RL: **Residue-residue potentials with a favorable contact pair term and an unfavorable hight packing density term, for simulation and threading.** *Journal of molecular biology* 1996, **256(3):**623-644.

14. Sippl MJ: **Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.** *Journal of computer-aided molecular design* 1993, **7:**473-501.
15. Solis AD, Rackovsky S: **Improvement of statistical potentials and threading score functions using information maximixation.** *Proteins* 2006, **62(4):**892-908.
16. Tozzini V: **Coarse-grained model for proteins.** *Current opinion in structural biology* 2005, **15:**144-150.
17. Seno F, Vendruscolo M, Maritan A, Banavar JR: **Optimal protein design procedures.** *Physical review letter* 1996, **77(9):**1901-1904.
18. Deutsch JM, Kurowski T: **New algorithm for protein design.** *Physical review letter* 1996, **76:**323-326.
19. Seno F, Micheletti M, Maritan A, Banavar JR: **Variational approach to protein design and extraction of interactional potentials.** *Physical review letter* 1998, **81:**2172-2175.
20. Rossi A, Maritan A, Micheletti C: **A novel iterative strategy for protein design.** *Journal of Chemical physics* 2000, **112(4):**2050-2055.
21. Rossi A, Micheletti C, Seno F, Maritan A: **A self-consistent knowledge-based approach to protein design.** *Biophysical journal* 2001, **80(1):**480-490.
22. Moult J: **Comparison of database potentials and molecular mechanics force fields.** *Current opinion in structural biology* 1997, **7(2):**194-1999.
23. Mendes J, Guerois R, Serrano L: **Energy estimation in protein design.** *Current opinion in structural biology* 2002, **12(4):**441-446.
24. Bowie JU, Luthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, **253(5016):**164-170.
25. Chiu TL, Goldstein RA: **Optimizing potentials for the inverse protein folding problem.** *Protein engineering* 1998, **11:**749-752.
26. Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N: **A maximum likelihood framework for protein design.** *BMC Bioinformatics* 2006, **7:**326-343.
27. Kuhlman B, Baker D: **Native protein sequences are close to optimal for their structure.** *PNAS* 2000, **97(19):**10383-10388.
28. Shakhnovich E, Gutin A: **Engineering of stable and fast-folding sequences of model proteins.** *Proceedings Natl Academy of sciences USA* 1993, **90(15):**7195-7199.
29. Sun S, Brem R, Chan R, Dill K: **Designing amino acid sequences to fold with good hydrophobic cores.** *Protein Engineering* 1995, **8(12):**1205-1213.
30. Pande VS, Grosberg AY, Tanaka T: **Statistical mechanics of simple model of protein folding and design.** *Biophysical journal* 1997, **73(6):**3192-3210.
31. Muse SV, Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to cholorplast genome.** *Mol Biol Evol* 1994, **11:**715-724.
32. Bastolla U, Porto M, Roman HE, Vendruscolo M: **A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank.** *BMC Evolutionary Biology* 2006, **6:**43.
33. Panjkovich A, Melo F, Marti-Renom M: **Evolutionary potentials: structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs.** *Genome Biology* 2008, **9:**R68.
34. Wang G, Dunbrack RLJ: **PISCES: a protein sequence culling server.** *Bioinformatics* 2003, **19(12):**1589-1591.
35. Hubbard SJ, Thornton JM: **NACCESS.** Computer Program, Department of Biochemistry and Molecular Biology, University College London; 1993.
36. Yang Z, Nielsen R, Goldman N, K P: **Codon substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155:**431-449.
37. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3. A Bayesian software for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25(17):**2286-2288.
38. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS: **Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenics.** *Science* 2001, **294(5550):**2348-2351.
39. Yang Z, Swanson WJ, Vacquier VD: **Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites.** *Molecular Biology and Evolution* 2000, **17:**1446-1455.
40. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Systematic Biology* 2003, **52(5):**696-704.
41. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *CABIOS* 1992, **8:**275-282.
42. Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites.** *Mol Biol Evol* 1993, **10:**1396-1401.

Appendix II

Supplementary material for chapter 3

Potential	*ΔCV*
Bfactor	0.056
ss (3 classes)	0.043
ss (10 classes)	0.066
torsion	0.120
solv	0.140
cont	0.221
dist	0.382
cont,solv	0.263
torsion,ss (3 classes)	0.139
torsion,ss (10 classes)	0.153
solv,Bfactor	0.153
solv,torsion	0.256
solv,Bfactor,torsion	0.270
dist,torsion	0.479
dist,solv	0.406
dist,solv,ss	0.443
dist,solv,torsion,ss	0.523
dist,solv,Bfactor,torsion,ss	0.535

Table S1: Cross validation scores for the different potentials obtained.

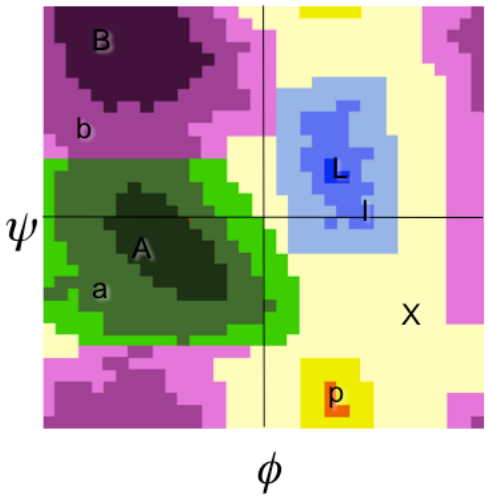


Figure S1: Specification of the discrete classes for backbone torsion angles. Ramachandran plot from Laskowski et al. (1996)

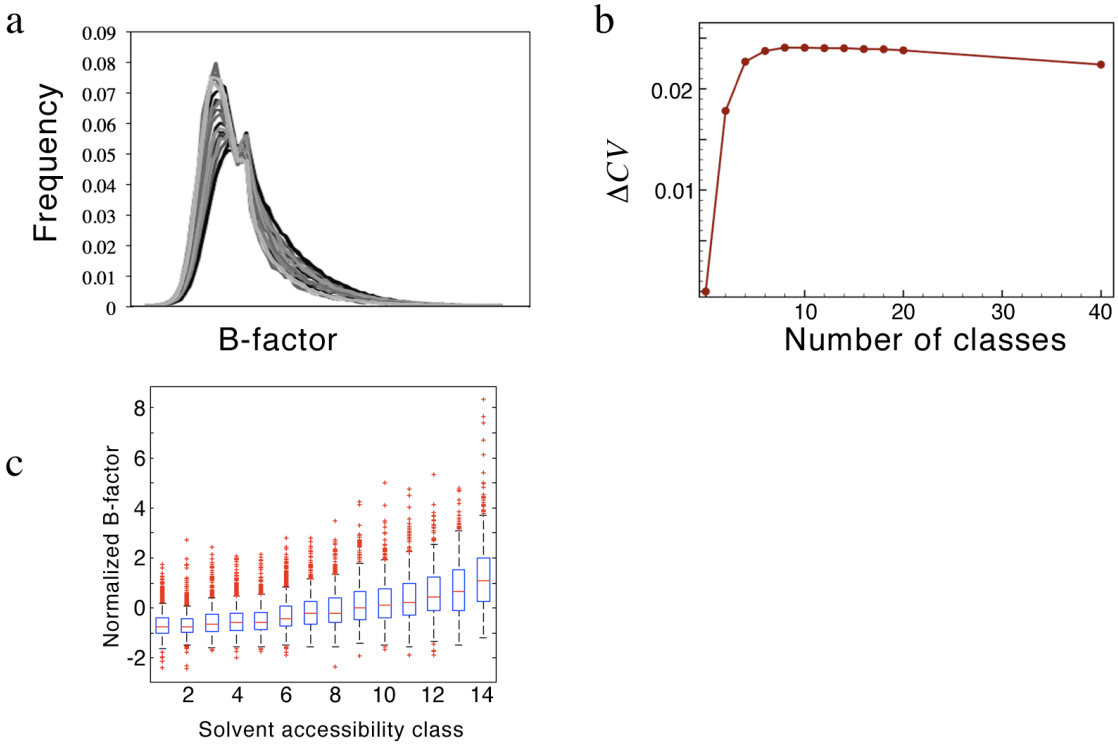


Figure S2: A) Distribution of B-factor for the different amino acids, in a nonredundant subset of PDB of 1,000 proteins. B-factor was calculated as the B-factor reported for the α-carbon of the residue, and normalized within each protein B) Evolution of cross-validation score of the potential as a function of the number of classes, for a potential based on α-carbon B-factor. C) Box plot of B-factor and solvent accessibility for 10,000 residues taken from nonredundant sets of PDB. The central mark in each box is the median; the edges of the boxes are the 25th and 75th percentiles; the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

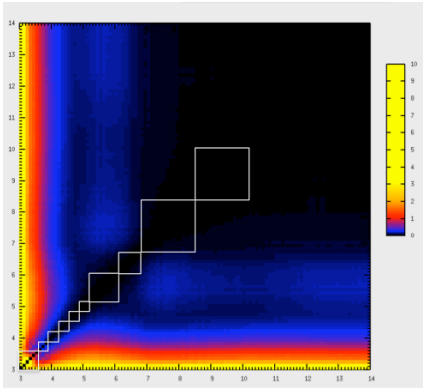
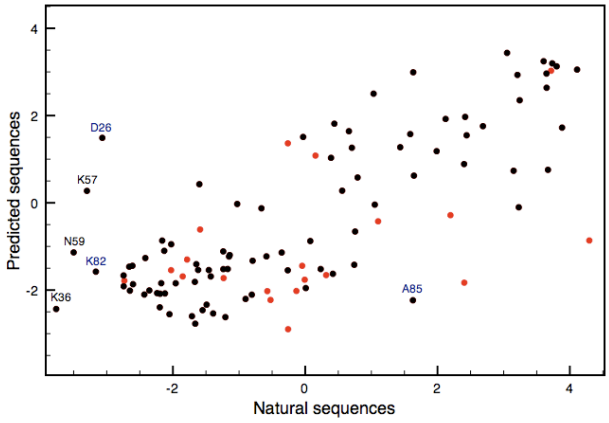


Figure S3: Pairwise comparison of the Kullback-Leibler divergence. The interval 0-25 Å was divided in bins of 0.25Å, and the distribution of pairwise interactions in each bin was compared to all the other bins, using the Küllback-Liebler divergence. The resulting values of divergence were represented graphically using the coloring scheme displayed, with black representing the lowest value, and yellow the highest. An example of the different definitions of distance classes tested is shown with white squares.

a



b

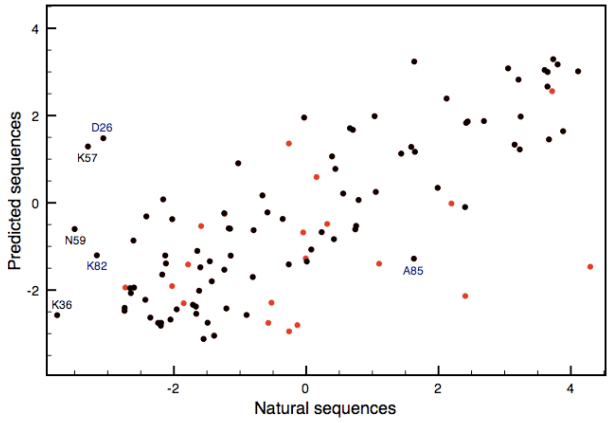
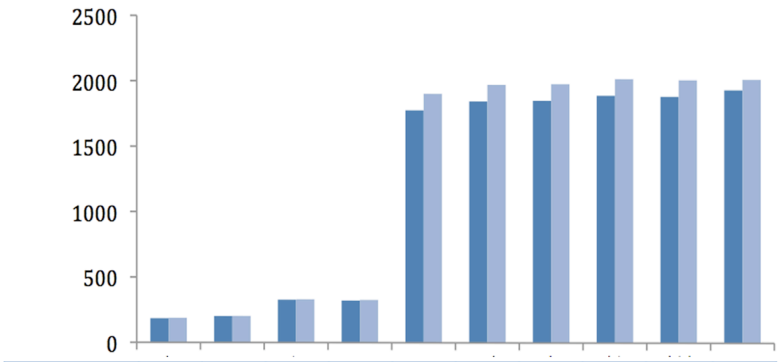


Figure S4: Comparison of hydrophobicity values of natural and predicted sequences, using the site-specific profiles from Figure 4. Each point represents the average hydrophobicity for a site *i* in natural sequences (X-axis) and in sequences designed (Y-axis) using the potentials a) $ML_{dist,solv}$, or b) $ML_{dist,solv,Bfactor,torsion}$. Red dots correspond to catalytic sites or sites in contact to ligand or metal. Some of the residues clearly deviating from the global tendency have been labeled with the position in the protein structure. Residues A85, D26 and K82 are in close contact; distance potentials fail to predict the polar interaction between D26 and K82, and instead predict a polar residue in position 85. Without considering these three residues or the functional sites (red dots), the correlation coefficients are a) 0.70 and b)0.71. Hydrophobicity was assigned using the scale of Kyte & Doolittle (1982).



ω					+	+	+	+	+	DP
dist	+	+	+	+		+	+	+	+	
solv		+	+	+			+	+	+	
torsion			+	+				+	+	
Bfactor				+					+	

Figure S5: Natural logarithm of Bayes factors for the models considered, applied to a dataset consisting of 34 eukaryotic sequences of Calmodulin. The two bars for each potential represent two duplicates of the numerical calculation. The first 4 columns correspond to the MG-SC model; the 5th column correspond to MG-NS; columns 6-9 correspond to the MG-NS-SC models; the last column corresponds to the MG-NS^{DP} model.

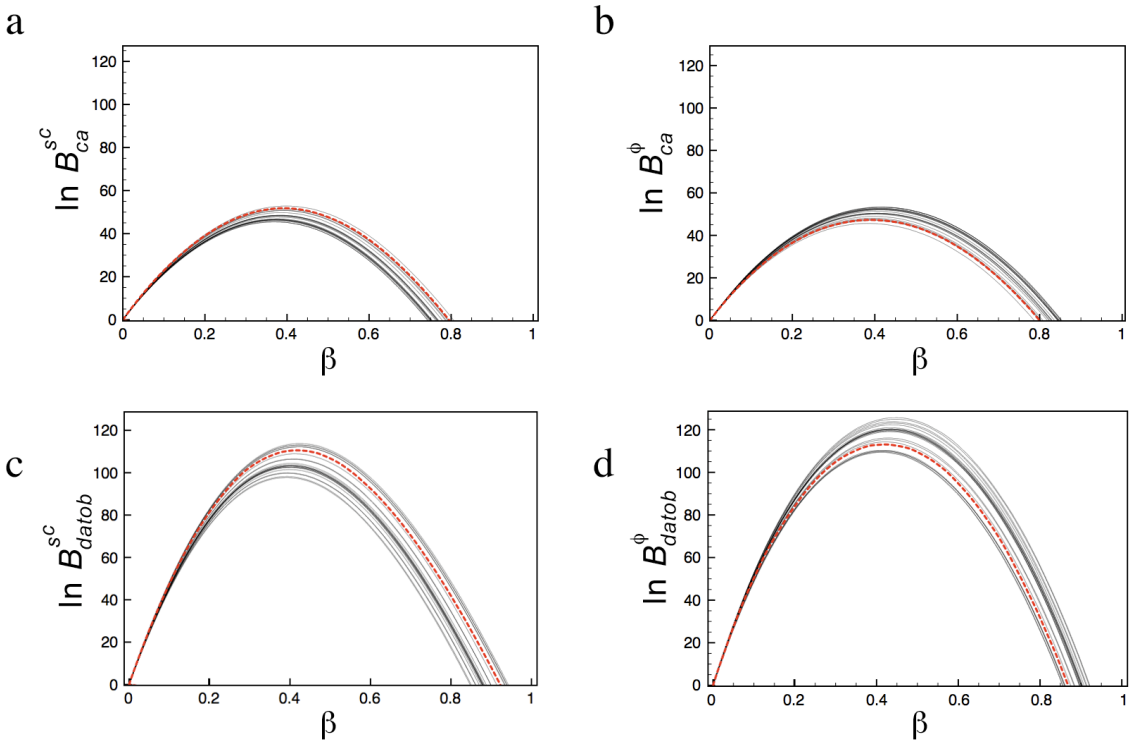


Figure S6: Trace plots representing the stationary factor $B_M^{s^c}$ (*a* and *c*) and the transient factor B_M^{ϕ} (*b* and *d*) as a function of β. The computation was performed on the ADH dataset, using the potentials $ML_{cont,solv}$ (*a* and *b*), or $ML_{dist,solv,Bfactor,torsion}$. (*c* and *d*). In each curve, a different sequence from the alignment is taken as s^c. The dashed line corresponds to the case where the native sequence is taken as s^c.

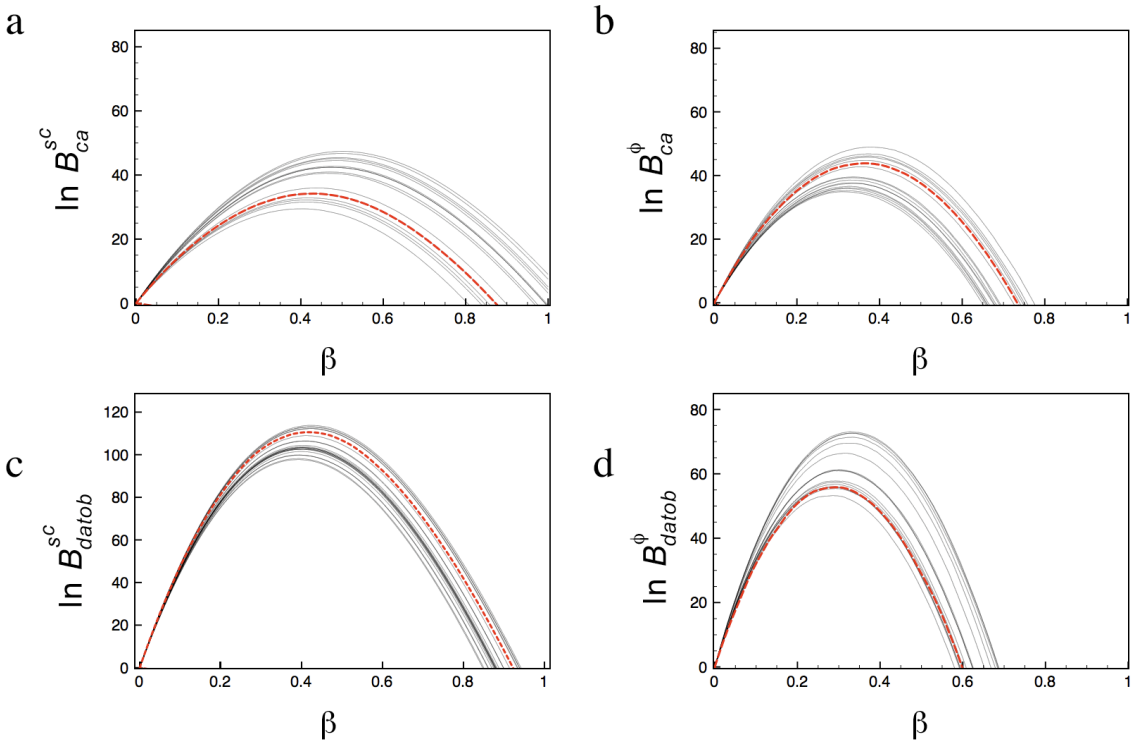


Figure S7: Trace plots representing the stationary factor $B_M^{s^c}$ (*a* and *c*) and the transient factor B_M^{ϕ} (*b* and *d*) as a function of β. The computation was performed on the β-globin dataset, using the potentials $ML_{cont,solv}$ (*a* and *b*), or $ML_{dist,solv,Bfactor,torsion}$. (*c* and *d*). In each curve, a different sequence from the alignment is taken as s^c. The dashed line corresponds to the case where the native sequence is taken as s^c.